

Target Detection Performance Bounds in Compressive Imaging*

Kalyani Krishnamurthy^{†1}, Rebecca Willett^{‡1} and Maxim Raginsky^{§2}

¹Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708.

²Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801.

August 15, 2012

Abstract

This paper describes computationally efficient approaches and associated theoretical performance guarantees for the detection of known targets and anomalies from few projection measurements of the underlying signals. The proposed approaches accommodate signals of different strengths contaminated by a colored Gaussian background, and perform detection without reconstructing the underlying signals from the observations. The theoretical performance bounds of the target detector highlight fundamental tradeoffs among the number of measurements collected, amount of background signal present, signal-to-noise ratio, and similarity among potential targets coming from a known dictionary. The anomaly detector is designed to control the number of false discoveries. The proposed approach does not depend on a known sparse representation of targets; rather, the theoretical performance bounds exploit the structure of a known dictionary of targets and the distance preservation property of the measurement matrix. Simulation experiments illustrate the practicality and effectiveness of the proposed approaches.

Keywords: Target Detection, Anomaly detection, False Discovery Rate, p-value, Incoherent Projections, Compressive sensing

*This work was supported by NSF Award No. DMS-08-11062, DARPA Grant No. HR0011-09-1-0036, and AFRL Grant No. FA8650-07-D-1221.

[†]kk63@duke.edu

[‡]willett@duke.edu

[§]maxim@illinois.edu

1 Introduction

The theory of compressive sensing (CS) has shown that it is possible to accurately *reconstruct* a sparse signal from few (relative to the signal dimension) projection measurements [9, 15]. Though such a reconstruction is crucial to visually inspect the signal, there are many instances where one is solely interested in identifying whether the underlying signal is one of several possible signals of interest. In such situations, a complete reconstruction is computationally expensive and does not optimize the correct performance metric. Recently, CS ideas have been exploited in [12, 16, 21] to perform target detection and classification from projection measurements, without reconstructing the underlying signal of interest. In [12, 21], the authors propose nearest-neighbor based methods to classify a signal $\mathbf{f} \in \mathbb{R}^N$ to one of m known signals given projection measurements of the form $\mathbf{y} = \mathbf{A}\mathbf{f} + \mathbf{n} \in \mathbb{R}^K$ for $K \leq N$, where $\mathbf{A} \in \mathbb{R}^{K \times N}$ is a known projection operator and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ is the additive Gaussian noise. This model is simple to analyze, but is impractical, since in reality, a signal is always corrupted by some kind of interference or background noise. Extension of the methods in [12, 21] to handle background noise is nontrivial. Though [16] provides a way to account for background contamination, it makes a strong assumption that the signal of interest and the background are sparse in bases that are incoherent. This might not always be true in many applications. Recent works on CS [2, 3] allow for the input signal \mathbf{f} to be corrupted by some pre-measurement noise $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma_b^2 \mathbf{I})$ such that one observes $\mathbf{y} = \mathbf{A}(\mathbf{f} + \mathbf{b}) + \mathbf{n}$, and study reconstruction performance as a function of the number of measurements, pre- and post-measurement noise statistics and the dimension of the input signal. In this work, however, we are interested in performing target detection without an intermediate reconstruction step. Furthermore, the increased utility of high-dimensional imaging techniques such as spectral imaging or videography in applications like remote sensing, biomedical imaging and astronomical imaging [20, 26, 35, 38–40, 47, 58] necessitates the extension of compressive target detection ideas to such imaging modalities to achieve reliable target detection from fewer measurements relative to the ambient signal dimensions.

For example, recent advances in compressive sensing (CS) have led to the development of new spectral imaging platforms which attempt to address challenges in conventional imaging platforms related to system size, resolution, and noise by acquiring fewer compressive measurements than spatio-spectral voxels [8, 14, 18, 50, 53, 57]. However, these system designs have a number of degrees of freedom which influence subsequent data analysis. For instance, the single-shot compressive spectral imager discussed in [18] collects *one* coded projection of each spectrum in the scene. One projection per spectrum is sufficient for reconstructing spatially homogeneous spectral images, since projections of neighboring locations can be combined to infer each spectrum. Significantly more projections are required for detecting targets of unknown strengths without the benefit of spatial homogeneity. We are interested in investigating how several such systems can be used in *parallel* to reliably detect spectral targets and anomalies from different coded projections.

In general, we consider a broadly applicable framework that allows us to account for background and sensor noise, and perform target detection directly from projection measurements of signals obtained at different spatial or temporal locations. The precise problem formulation is provided below.

1.1 Problem formulation

Let us assume access to a dictionary of possible targets of interest $\mathcal{D} = \{\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \dots, \mathbf{f}^{(m)}\}$, where $\mathbf{f}^{(j)} \in \mathbb{R}^N$ for $j = 1, \dots, m$ is unit-norm. Our measurements are of the form

$$\mathbf{z}_i = \Phi(\alpha_i \mathbf{f}_i^* + \mathbf{b}_i) + \mathbf{w}_i \quad (1)$$

where

- $i \in \{1, \dots, M\}$ indexes the spatial or temporal locations at which data are collected;
- $\alpha_i \geq 0$ is a measure of the signal-to-noise ratio at location i , which is either known or estimated from observations;
- $\Phi \in \mathbb{R}^{K \times N}$ for $K < N$, is a measurement matrix to be specified in Sec. 2;
- $\mathbf{b}_i \in \mathbb{R}^N \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$ is the background noise vector, and $\mathbf{w}_i \in \mathbb{R}^K \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ is the i.i.d. sensor noise.

For example, in the case of spectral imaging \mathbf{f}_i^* represents the spectrum at the i^{th} spatial location, and in video sequences \mathbf{f}_i^* represents the vectorized image frame obtained at the i^{th} time interval. In this paper we consider the following target detection problems:

1. Dictionary signal detection (DSD): Here we assume that each $\mathbf{f}_i^* \in \mathcal{D}$ for $i \in \{1, \dots, M\}$, and our task is to detect all instances of one target signal $\mathbf{f}^{(j)} \in \mathcal{D}$ for some unknown $j \in \{1, \dots, m\}$, *i.e.*, to locate $S = \{i : \mathbf{f}_i^* = \mathbf{f}^{(j)}\}$. DSD is useful in contexts in which we know the makeup of a scene and wish to focus our attention on the locations of a particular signal. For instance, in spectral imaging, DSD is used to study a scene of interest by classifying every spectrum in the scene to different known classes [36, 38]. In a video setup, DSD could be used to classify video segments to one of several categories (such as news, weather, sports, etc.) by projecting the video sequence to an appropriate feature space and comparing the feature vectors to the ones in a known dictionary [55].
2. Anomalous signal detection (ASD): Here, our task is to detect all signals which are *not* members of our dictionary, *i.e.*, detect $S = \{i : \mathbf{f}_i^* \notin \mathcal{D}\}$. (This is akin to anomaly detection methods in the literature which are based on nominal, nonanomalous training samples [23, 46].) For instance, ASD may be used when we know most components of a spectral image and wish to identify all spectra which deviate from this model [45].

Our goal is to accurately perform DSD or ASD without reconstructing the spectral input \mathbf{f}_i^* from \mathbf{z}_i for $i \in \{1, \dots, M\}$. Accounting for background is a crucial issue. Typically, the background corresponding to the scene of interest and the sensor noise are modeled together by a colored multivariate Gaussian distribution [37]. However, in our case, it is important to distinguish the two because of the presence of the projection operator Φ . The projection operator acts upon the background spectrum in the same way as on the target spectrum, but it does not affect the sensor noise. We assume that \mathbf{b}_i and \mathbf{w}_i are independent of each other, and the prior probabilities of different targets in the dictionary $p^{(j)} = \mathbb{P}(\mathbf{f}_i^* = \mathbf{f}^{(j)})$ for $j \in \{1, \dots, m\}$ are known in advance. If these probabilities are unknown, then the targets can be considered equally likely. Given this setup, our goal is to develop suitable target and anomaly detection approaches, and provide theoretical guarantees on their performances.

In this paper we develop detection performance bounds which show how performance scales with the number of detectors in a compressive setting as a function of SNR, the similarity between potential targets in a known dictionary, and their prior probabilities. Our bounds are based on a detection strategy which operates directly on the collected data as opposed to first reconstructing each \mathbf{f}_i^* and then performing detection on the estimated signals. Reconstruction as an intermediate step in detection may be appealing to end users who wish to visually inspect spectral images instead of relying entirely on an automatic detection algorithm. However, using this intermediate step has two potential pitfalls. First, the Rao–Blackwell theorem [6] tells us that an optimal detection algorithm operating on the processed data (*i.e.*, not sufficient statistics) *cannot* perform better than an optimal detection algorithm operating on the raw data. In other words, optimal performance is possible on the raw data, but we have no such performance guarantee for the reconstructed signals. Second, the relationship between reconstruction errors and detection performance is not well understood in many settings. Although we do not reconstruct the underlying signals, our performance bounds are intimately related to the signal resolution needed to achieve the signal diversity present in our dictionary. Since we have many fewer observations than the signals at this resolution, we adopt the “compressive” terminology.

1.2 Performance metric

To assess the performance of our detection strategies, we consider the False Discovery Rate (FDR) metric and related quantities developed for multiple hypothesis testing problems [5]. Since we collect M independent observations of potentially different signals, we are simultaneously conducting M hypothesis tests when we search for targets. Unlike the probability of false alarm, which measures the probability of falsely declaring a target for a *single* test, the FDR measures the *fraction* of declared targets that are false alarms, that is, it provides information about the entire set of M hypotheses instead of just one. More

formally, the FDR is given by,

$$\text{FDR} = \mathbb{E} \left[\frac{V}{R} \right],$$

where V is the number of falsely rejected null hypotheses, and R is the total number of rejected null hypotheses. Controlling the false discovery rate in a multiple hypothesis testing framework is akin to designing a constant false alarm rate (CFAR) detector in spectral target detection applications that keeps the false alarm rate at a desired level irrespective of the background interference and sensor noise statistics [36].

1.3 Previous investigations

Much of the classical target detection literature [24, 27, 29, 34, 43] assume that each target lies in a P -dimensional subspace of \mathbb{R}^N for $P < N$. The subspace in which the target lies is often assumed to be known or specified by the user, and the variability of the background is modeled using a probability distribution. Given knowledge of the target subspace, background statistics and sensor noise statistics, detection methods based on LRTs (likelihood ratio tests) and GLRTs (generalized likelihood ratio tests) have been proposed in [24, 27, 29, 34, 43, 44]. A subspace model is optimal if the subspace in which targets lie is known in advance. However, in many applications, such subspaces might be hard to characterize. An alternative, and a more flexible option is to assume that the high-dimensional target exhibits some low-dimensional structure that can be exploited to perform efficient target detection. This approach is utilized in this work and in [21] where the target signal in \mathbb{R}^N is assumed to come from a dictionary of m known signals such that $m \ll N$, and in [12], where the targets are assumed to lie in a low-dimensional manifold embedded in high-dimensional target space.

Recently, several methods for target or anomaly detection that rely on recovering the full spatio-spectral data from projection measurements [41, 56] have been proposed. However, they are computationally intensive and the detection performance associated with these reconstructions is unknown. Other researchers have exploited compressive sensing to perform target detection and classification without reconstructing the underlying signal [12, 16, 21]. In [16], the authors propose a matching pursuit based algorithm, called the *Incoherent Detection and Estimation Algorithm* (IDEA), to detect the presence of a signal of interest against a strong interfering signal from noisy projection measurements. The algorithm is shown to perform well on experimental data sets under some strong assumptions on the sparsity of the signal of interest and the interfering signal. In [12], the authors develop a classification algorithm called the *smashed filter* to classify an image in \mathbb{R}^N to one of m known classes from K projections of the signal, where $K < N$. The underlying image is assumed to lie on a low-dimensional manifold, and the algorithm finds the closest match from the m known classes by performing a nearest neighbor search over the m different manifolds. The projection measurements are chosen to preserve the distances

among the manifolds. Though [12] offers theoretical bounds on the number of measurements necessary to preserve distances among different manifolds, it is not clear how the performance scales with K or how to incorporate background models into this setup. Moreover, this approach may be computationally intensive since it involves learning and searching over different manifolds. In [21], the authors use a nearest-neighbor classifier to classify an N -dimensional signal to one of m equally likely target classes based on $K < N$ random projections, and provide theoretical guarantees on the detector performance. While the method discussed in [21] is computationally efficient, it is nontrivial to extend to the case of target detection with colored background noise and nonequiprobable targets. Furthermore, their performance guarantees cannot be directly extended to our problem since we focus on error measures that let us analyze the performance of multiple hypothesis tests *simultaneously* as opposed to the above methods that consider compressive classification performance for a *single* hypothesis test.

The authors of a more recent work [17] extend the classical RX anomaly detector [42] to *directly* detect anomalies from random, orthonormal projection measurements without an intermediate reconstruction step. They numerically show how the detection probability improves as a function of the signal-to-noise ratio when the number of measurements changes. Though probability of detection is a good performance measure, in many applications controlling the false discoveries below a desired level is more crucial. As a result, in our work, we propose an anomaly detection method that controls the false discovery rate below a desired level.

1.4 Contributions

This paper makes the following contributions to the above literature:

- A *compressive target detection* approach, which (a) is computationally efficient, (b) allows for the signal strengths of the targets to vary with spatial location, (c) allows for backgrounds mixed with potential targets, (d) considers targets with different a priori probabilities, and (e) yields theoretical guarantees on detector performance. This paper unifies preliminary work by the authors [30, 31], presents previously unpublished aspects of the proofs, and contains updated experimental results.
- A computationally efficient *anomaly detection* method that detects anomalies of different strengths from projection measurements and also controls the false discovery rate at a desired level.
- A *whitening filter approach* to compressive measurements of signals with background contamination, and associated analysis leading to bounds on the amount of background to which our detection procedure is robust.

The above theoretical results, which are the main focus of this paper, are supported with simulation studies in Sec. 6. Classical detection methods described in [7, 19, 24, 27, 29, 33, 34, 36, 37, 42–45, 49] do not establish performance bounds

as a function of signal resolution or target dictionary properties and rely on relatively *direct observation* models which we show to be suboptimal when the detector size is limited. The methods in [12] and [16] do not contain performance analysis, and our analysis builds upon the analysis in [21] to account for several specific aspects of the compressive target detection problem.

2 Whitening compressive observations

Before we present our detection methods for DSD and ASD problems respectively, we briefly discuss a whitening step that is common to both our problems of interest.

Let us suppose that there are enough background training data available to estimate the background mean $\boldsymbol{\mu}_b$ and covariance matrix $\boldsymbol{\Sigma}_b$. We can assume without loss of generality that $\boldsymbol{\mu}_b = \mathbf{0}$ since $\boldsymbol{\Phi}\boldsymbol{\mu}_b$ can be subtracted from \mathbf{y} . Given the knowledge of the background statistics, we can transform the background and sensor noise model $\boldsymbol{\Phi}\mathbf{b}_i + \mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi}\boldsymbol{\Sigma}_b\boldsymbol{\Phi}^T + \sigma^2\mathbf{I})$ discussed in (1) to a simple white Gaussian noise model by multiplying the observations \mathbf{z}_i , $i \in \{1, \dots, M\}$, by the *whitening filter* $C_{\boldsymbol{\Phi}} \triangleq (\boldsymbol{\Phi}\boldsymbol{\Sigma}_b\boldsymbol{\Phi}^T + \sigma^2\mathbf{I})^{-1/2}$. This whitening transformation reduces the observation model in (1) to

$$\mathbf{y}_i = C_{\boldsymbol{\Phi}} \underbrace{(\boldsymbol{\Phi}(\alpha_i \mathbf{f}_i^* + \mathbf{b}_i) + \mathbf{w}_i)}_{\mathbf{z}_i} = \alpha_i \mathbf{A} \mathbf{f}_i^* + \mathbf{n}_i \quad (2)$$

where

$$\mathbf{A} = C_{\boldsymbol{\Phi}} \boldsymbol{\Phi}, \quad (3)$$

and $\mathbf{n}_i = C_{\boldsymbol{\Phi}}(\boldsymbol{\Phi}\mathbf{b}_i + \mathbf{w}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. To verify that $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, observe that

$$\mathbf{n}_i = C_{\boldsymbol{\Phi}}(\boldsymbol{\Phi}\mathbf{b}_i + \mathbf{w}_i) \sim \mathcal{N}\left(\mathbf{0}, \underbrace{C_{\boldsymbol{\Phi}}(\boldsymbol{\Phi}\boldsymbol{\Sigma}_b\boldsymbol{\Phi}^T + \sigma^2\mathbf{I})C_{\boldsymbol{\Phi}}^T}_{\mathbf{I}}\right).$$

We can now choose $\boldsymbol{\Phi}$ so that the corresponding \mathbf{A} has certain desirable properties as detailed in Sec. 3 and Sec. 5.

For a given \mathbf{A} , the following theorem provides a construction of $\boldsymbol{\Phi}$ that satisfies (3) and a bound on the maximum tolerable background contamination:

Theorem 1. *Let $\mathbf{B} = \mathbf{I} - \mathbf{A}\boldsymbol{\Sigma}_b\mathbf{A}^T$. If the largest eigenvalue of $\boldsymbol{\Sigma}_b$ satisfies*

$$\lambda_{\max} < \frac{1}{\|\mathbf{A}\|^2}, \quad (4)$$

where $\|\mathbf{A}\|$ is the spectral norm of \mathbf{A} , then \mathbf{B} is positive definite and $\boldsymbol{\Phi} = \sigma\mathbf{B}^{-1/2}\mathbf{A}$ is a sensing matrix, which can be used in conjunction with a whitening filter to produce observations modeled in (2).

The proof of this theorem is provided in Appendix A. This theorem draws an interesting relationship between the maximum background perturbation that the system can tolerate and the spectral norm of the measurement matrix, which in turn varies with K and N . Hardware designs such as those in [14, 50] use spatial light modulators and digital micro mirrors, which allow the measurement matrix Φ to be adjusted easily in response to changing background statistics and other operating conditions.

In the sections that follow, we consider collecting measurements of the form $\mathbf{y}_i = \alpha_i \mathbf{A} \mathbf{f}_i^* + \mathbf{n}_i$ given in (2), where \mathbf{f}_i^* is the target of interest for $i = 1, \dots, M$, and $\mathbf{A} \in \mathbb{R}^{K \times N}$ is a sensing matrix that satisfies (3). It is assumed that any background contamination has been eliminated with the whitening procedure described in this section.

3 Dictionary signal detection

Suppose that the end user wants to test for the presence of one known target versus the rest, but it is not known a priori which target from \mathcal{D} the user wants to detect. In this case, let us cast the DSD problem as a multiple hypothesis testing problem of the form

$$\mathcal{H}_{0i}^{(j)} : \mathbf{f}_i^* = \mathbf{f}^{(j)} \quad \text{vs.} \quad \mathcal{H}_{1i}^{(j)} : \mathbf{f}_i^* \neq \mathbf{f}^{(j)} \quad (5)$$

where $\mathbf{f}^{(j)} \in \mathcal{D}$ is the target of interest and $i = 1, \dots, M$.

3.1 Decision rule

We define our decision rule corresponding to target $\mathbf{f}^{(j)} \in \mathcal{D}$ in terms of a set of significance regions $\Gamma_i^{(j)}$ such that one rejects the i^{th} null hypothesis if its test statistic \mathbf{y}_i falls in the i^{th} significance region. Specifically, $\Gamma_i^{(j)}$ is defined according to

$$\Gamma_i^{(j)} = \left\{ \mathbf{y} : \log \mathbb{P}(\mathbf{f}_i^* = \mathbf{f}^{(j)} | \mathbf{y}_i, \alpha_i, \mathbf{A}) \leq \log \mathbb{P}(\mathbf{f}_i^* = \mathbf{f}^{(\ell)} | \mathbf{y}_i, \alpha_i, \mathbf{A}) \text{ for some } \ell \in \{1, \dots, m\}, \ell \neq j \right\}, \quad (6)$$

where $\log \mathbb{P}(\mathbf{f}_i^* = \mathbf{f}^{(j)} | \mathbf{y}_i, \alpha_i, \mathbf{A}) = \frac{K}{2} \log \left(\frac{1}{2\pi} \right) - \frac{\|\mathbf{y}_i - \alpha_i \mathbf{A} \mathbf{f}^{(j)}\|^2}{2} + \log p^{(j)}$ is the logarithm of the a posteriori probability density of the target $\mathbf{f}^{(j)}$ at the i^{th} spatial location given the observations \mathbf{y}_i , the signal-to-noise ratio α_i and the sensing matrix \mathbf{A} , and $p^{(j)}$ is the a priori probability of target class j . Note that the process of determining these decision regions involves a sequence of nearest-neighbor calculations, so the computational complexity scales with the number of classes m . In this work, we operate under the assumption that m is much smaller than the dimensionality of the datasets we consider. For example, if we consider spectral images, then the number of objects (signal classes) that make up a scene of interest is often smaller than the number of voxels in the

image. This assumption is not unrealistic and has been exploited in earlier work such as [36] and the references therein. In most of the prior work we have surveyed [10, 11], the number of signal classes is less than 35, which doesn't make our approach intractable.

The decision rule can be formally expressed in terms of the significance regions as follows:

$$\text{reject } \mathcal{H}_{0i}^{(j)} \text{ if the test statistic } \mathbf{y}_i \in \Gamma_i^{(j)}. \quad (7)$$

We analyze this detector by extending the positive False Discovery Rate (pFDR) error measure introduced by Storey to characterize the errors encountered in performing multiple, independent and *nonidentical* hypothesis tests simultaneously [48]. The pFDR, discussed formally below, is the fraction of falsely rejected null hypotheses among the total number of rejected null hypotheses, subject to the positivity condition that one rejects at least one null hypothesis. The pFDR is similar to the FDR except that the positivity condition is enforced here. In our context, the positivity condition means that we declare at least one signal to be a nontarget, which in turn implies that the scene of interest is composed of more than one object in the case of spectral imaging, or that the scene is not static in the case of video imaging.

Consider a collection of significance regions $\mathbf{\Gamma} = \{\Gamma_i^{(j)} : i = 1, \dots, M\}$, such that one declares $\mathcal{H}_{1i}^{(j)}$ if the test statistic $\mathbf{y}_i \in \Gamma_i^{(j)}$. The pFDR for multiple, *nonidentical* hypothesis tests can be defined in terms of the significance regions as follows:

$$\text{pFDR}^{(j)}(\mathbf{\Gamma}) = \mathbb{E} \left[\frac{V(\mathbf{\Gamma})}{R(\mathbf{\Gamma})} \middle| R(\mathbf{\Gamma}) > 0 \right] \quad (8)$$

where

$$V(\mathbf{\Gamma}) = \sum_{i=1}^M \mathbb{I}_{\{\mathbf{y}_i \in \Gamma_i^{(j)}\}} \mathbb{I}_{\{\mathcal{H}_{0i}\}} \quad (9)$$

is the number of falsely rejected null hypotheses,

$$R(\mathbf{\Gamma}) = \sum_{i=1}^M \mathbb{I}_{\{\mathbf{y}_i \in \Gamma_i^{(j)}\}} \quad (10)$$

is the total number of rejected null hypotheses, and $\mathbb{I}_{\{E\}} = 1$ if event E is true and 0 otherwise. In our setup, the pFDR corresponds to the expected ratio of the number of missed targets to the number of signals declared to be nontargets subject to the condition that at least one signal is declared to be a nontarget. (Note that this ratio is traditionally referred to as the positive false *nondiscovery* rate (pFNR), but is technically the pFDR in this context because of our definitions of the null and alternate hypotheses.) The theorem below presents our main result:

Theorem 2. *Given observations of the form (2), if one performs multiple, independent, nonidentical hypothesis tests of the form (5) and decides according to (7), then the worst-case pFDR given by $\text{pFDR}_{\max} = \max_{j \in \{1, \dots, m\}} \text{pFDR}^{(j)}(\mathbf{\Gamma})$, satisfies the following bound:*

$$\text{pFDR}_{\max} \leq \min \left(1, \frac{(\text{P}_e)_{\max}}{1 - p_{\max} - (\text{P}_e)_{\max}} \right) \quad (11)$$

where

$$\begin{aligned} p_{\max} &= \max_{j \in \{1, \dots, m\}} p^{(j)}, \\ (\text{P}_e)_{\max} &= \max_{i \in \{1, \dots, M\}} \mathbb{P} \left(\hat{\mathbf{f}}_i \neq \mathbf{f}_i^* \right), \text{ and} \\ \hat{\mathbf{f}}_i &= \arg \max_{\mathbf{f} \in \mathcal{D}} \mathbb{P} \left(\mathbf{f}_i^* = \mathbf{f} \mid \mathbf{y}_i, \alpha_i, \mathbf{A} \right). \end{aligned} \quad (12)$$

The proof of this theorem is detailed in Appendix B. A key element of our proof is the adaptation of the techniques from [48] to *nonidentical* independent hypothesis tests.

3.2 An achievable bound on the worst-case pFDR

Theorem 2 in the preceding section shows that, for a given \mathbf{A} , the worst-case pFDR is bounded from above by a function of the worst-case misclassification probability. In this section, we use this theorem to establish an achievable bound on the worst-case pFDR that explicitly depends on the number of observations K , signal strengths $\{\alpha_i\}_{i=1}^M$, similarity among different targets of interest, and a priori target probabilities.

Let us first define the quantities

$$\begin{aligned} d_{\min} &= \min_{\mathbf{f}^{(i)}, \mathbf{f}^{(j)} \in \mathcal{D}, i \neq j} \|\mathbf{f}^{(i)} - \mathbf{f}^{(j)}\| \\ p_{\min} &= \min_{j \in \{1, \dots, m\}} p^{(j)} \\ \alpha_{\min} &= \min_{i \in \{1, \dots, M\}} \alpha_i. \end{aligned}$$

Then we have the following theorem, whose proof is given in Appendix C:

Theorem 3. *Let λ_{\max} denote the largest eigenvalue of $\mathbf{\Sigma}_b$. For a given $0 < \epsilon < 1 - p_{\max}$, assume that K and N are sufficiently large so that the following*

conditions hold:

$$1 - p_{\max} - \epsilon \geq \frac{1 - p_{\min}}{p_{\min}} \left(1 + \frac{\alpha_{\min}^2 d_{\min}^2}{4K\sigma^2} \right)^{-\frac{K}{2}} + 2 \exp \left(-\frac{(K+N)\epsilon^2}{2} \right) \quad (13a)$$

$$\lambda_{\max} < \frac{1}{(1+\epsilon)^2 \left(\sqrt{\frac{N}{K}} + 1 \right)^2}, \quad (13b)$$

$$K > \frac{2 \log \left(\frac{2}{p_{\min}} \frac{1-p_{\min}}{1-p_{\max}} \right)}{\log \left(1 + \frac{\alpha_{\min}^2 d_{\min}^2}{4K} \right)}. \quad (13c)$$

Then there exists a $K \times N$ sensing matrix \mathbf{A} that satisfies the condition of Theorem 1, and for which

$$\begin{aligned} \text{pFDR}_{\max} &\leq \frac{1}{p_{\min}} \left(\frac{1 - p_{\max}}{1 - p_{\min}} \left(1 + \frac{\alpha_{\min}^2 d_{\min}^2}{4K} \right)^{\frac{K}{2}} - \frac{1}{p_{\min}} \right)^{-1} + \\ &\quad \frac{2(1 - p_{\max})}{\epsilon^2} \exp \left(-\frac{(K+N)\epsilon^2}{2} \right). \end{aligned} \quad (14)$$

This result has the following implications and consequences:

1. For a given N , the upper bound (13b) on λ_{\max} increases as K increases, which implies that the system can tolerate more background perturbation if we collect more measurements.
2. The pFDR bound (14) decays with the increase in the values of K , d_{\min} and α_{\min} , and increases as p_{\min} decreases. For a fixed p_{\max} , p_{\min} , α_{\min} and d_{\min} , the bound in (14) enables one to choose a value of K to guarantee a desired pFDR value.
3. The dominant part of the bound (14) is independent of N , and is only a function of K , p_{\max} , p_{\min} , α_{\min} , and d_{\min} . The lack of dependence on N is not unexpected. Indeed, when we are interested in preserving pairwise distances among the members of a fixed dictionary of size m , the Johnson–Lindenstrauss lemma [25] says that, with high probability, $K = \mathcal{O}(\log m)$ random Gaussian projections suffice, regardless of the ambient dimension N . This is precisely the regime we are working with here.
4. The bound on K given in (13c) increases logarithmically with the increase in the difference between p_{\max} and p_{\min} . This is to be expected since one would need more measurements to detect a less probable target as our decision rule weights each target by its a priori probability. If all targets are equally likely, then $p_{\max} = p_{\min} = 1/m$, and $K = \mathcal{O}(\log m)$ is sufficient provided $\alpha_{\min}^2 d_{\min}^2$ is sufficiently large such that

$$\log \left(1 + \frac{\alpha_{\min}^2 d_{\min}^2}{4K} \right) > \log \left(1 + \frac{\alpha_{\min}^2 d_{\min}^2}{4N} \right) > 1$$

(where the first inequality holds since $K < N$). In addition, the lower bound on K also illustrates the interplay between the signal strength of the targets, the similarity among different targets in \mathcal{D} , and the number of measurements collected. A small value of d_{\min} suggests that the targets in \mathcal{D} are very similar to each other, and thus α_{\min} and K need to be high enough so that similar targets can still be distinguished. The experimental results discussed in Sec. 6 illustrate the tightness of the theoretical results discussed here.

Inspection of the proof shows that if \mathbf{A} is generated according to a Gaussian distribution, then the conditions of Theorem 3 will be met with high probability.

4 Extension to a manifold-based target detection framework

The DSD problem formulation in Sec. 1.1 is accurate if the signals in the dictionary are faithful representations of the target signals that we observe. In reality, however, the target signals will differ from the dictionary signals owing to the differences in the experimental conditions under which they are collected. For instance, in spectral imaging applications, the observed spectrum of any material will not match the reference spectrum of the same material observed in a laboratory because of the differences in atmospheric and illumination conditions. To overcome this problem, one could form a large dictionary to account for such uncertainties in the target signals and perform target detection according to the approaches discussed in Sec. 2 and Sec. 3. A potential drawback with this approach is that our theoretical performance bound increases with the size of \mathcal{D} through p_{\min} and d_{\min} . Instead, one could reasonably model the target signals observed under different experimental conditions to lie in a low-dimensional submanifold of the high-dimensional ambient signal space as shown to be true for spectral images in [22]. We can exploit this result to extend our analysis to a much broader framework that accounts for uncertainties in our dictionary.

Let us consider a dictionary of manifolds $\mathcal{D}_{\mathcal{M}} = \{\mathcal{M}^{(1)}, \dots, \mathcal{M}^{(m)}\}$ corresponding to m different target classes, and that \mathbf{f}_i^* for $i \in \{1, \dots, M\}$ is in one of the manifolds in $\mathcal{D}_{\mathcal{M}}$. Considering an observation model of the form given in (2), our goal is to determine $\{i : \mathbf{f}_i^* \in \mathcal{M}^{(j)}\}$, where $j \in \{1, \dots, m\}$ is the target class of interest. Let us assume that all target classes are equally likely to keep the presentation simple, though the analysis extends to the case where the targets classes have different a priori probabilities. Suppose that we collect independent sets of measurements $\{\mathbf{y}_i\}_{i=1}^M$ and $\{\tilde{\mathbf{y}}_i\}_{i=1}^M$. Then, we can use the following two-step procedure to extend our DSD method to this manifold-based framework:

1. Given $\{\mathbf{y}_i\}$, form a data-dependent dictionary $\mathcal{D}_{\mathbf{y}_i} = \{\tilde{\mathbf{f}}_i^{(1)}, \dots, \tilde{\mathbf{f}}_i^{(m)}\}$ corresponding to each \mathbf{y}_i by finding its nearest-neighbor in each manifold:

$$\tilde{\mathbf{f}}_i^{(\ell)} = \arg \max_{\mathbf{f} \in \mathcal{M}^{(\ell)}} \mathbb{P}(\mathbf{y}_i | \mathbf{f}_i^* = \mathbf{f}, \alpha_i, \mathbf{A})$$

for $\ell \in \{1, \dots, m\}$ and $i = 1, \dots, M$.

2. Given $\{\tilde{\mathbf{y}}_i\}$ and corresponding $\{\mathcal{D}_{\mathbf{y}_i}\}$, find

$$\hat{\mathbf{f}}_i = \arg \max_{\tilde{\mathbf{f}} \in \mathcal{D}_{\mathbf{y}_i}} \mathbb{P} \left(\tilde{\mathbf{y}}_i | \mathbf{f}_i^* = \tilde{\mathbf{f}}, \alpha_i, \mathbf{A} \right)$$

and declare that the i^{th} observed spectrum corresponds to class j if $\hat{\mathbf{f}}_i = \tilde{\mathbf{f}}_i^{(j)}$.

This two-step procedure is studied in [12] for the case $\{\mathbf{y}_i\} = \{\tilde{\mathbf{y}}_i\}$ where the authors provide bounds on the number of projection measurements needed to preserve distances among manifolds. However, they do not offer associated target detection performance guarantees. Our analysis and the theoretical performance bounds extend directly to this framework if we collect two sets of observations as discussed above. Specifically, the hypothesis tests corresponding to the second step can be written as

$$\mathcal{H}_{0i} : \mathbf{f}_i^* = \tilde{\mathbf{f}}_i^{(j)} \text{ vs. } \mathcal{H}_{1i} : \mathbf{f}_i^* \neq \tilde{\mathbf{f}}_i^{(j)}$$

where $\tilde{\mathbf{f}}_i^{(j)} \in \mathcal{D}_{\mathbf{y}_i}$ for $i = 1, \dots, M$. Since the dictionary in this case changes with i , these tests are *nonidentical*. This is another instance where our extension of pFDR-based analysis towards simultaneous testing of multiple, independent, and nonidentical hypothesis tests (8) is very significant. Following the proof techniques discussed in the appendix, we can straightforwardly show that the bound in (14) in this manifold setting holds with $p_{\min} = p_{\max} = 1/m$ since all target classes are assumed to be equally likely here, and $d_{\min} = \min_{i \in \{1, \dots, M\}} d_i$ where

$$d_i = \min_{\tilde{\mathbf{f}}_i^{(\ell)}, \tilde{\mathbf{f}}_i^{(k)} \in \mathcal{D}_{\mathbf{y}_i}, \ell \neq k} \|\tilde{\mathbf{f}}_i^{(\ell)} - \tilde{\mathbf{f}}_i^{(k)}\|.$$

5 Anomalous signal detection

The target detection approach discussed above assumes that the target signal of interest resides in a dictionary that is available to the user. However, in some applications (such as military applications and surveillance), one might be interested in detecting objects *not* in the dictionary. In other words, the target signals of interest are anomalous and are not available to the user. In this section we show how the target detection methods discussed above can be extended to anomaly detection. In particular, we exploit the distance preservation property of the sensing matrix \mathbf{A} to detect anomalous targets from projection measurements.

5.1 Problem formulation

Given observations of the form in (2), we are interested in detecting whether $\mathbf{f}^* \in \mathcal{D}$ or \mathbf{f}^* is anomalous. Let us write the anomaly detection problem as the

following multiple hypothesis test:

$$\mathcal{H}_{0i} : \|\mathbf{f}_i^* - \mathbf{f}\| \leq \tau \text{ for some } \mathbf{f} \in \mathcal{D} \quad (15a)$$

$$\mathcal{H}_{1i} : \|\mathbf{f}_i^* - \mathbf{f}\| > \tau \text{ for all } \mathbf{f} \in \mathcal{D} \quad (15b)$$

where $\tau \in [0, \sqrt{2}]$ is a user-defined threshold that encapsulates our uncertainty about the accuracy with which we know the dictionary.¹ In particular, τ controls how different a signal needs to be from every dictionary element to truly be considered anomalous. In the absence of any prior knowledge on the targets of interest, τ can simply be set to zero. The null hypothesis in this setting models the normal behavior, while the alternative hypothesis models the abnormal or anomalous behavior. This formulation is consistent with the literature [17, 45].

Note that the definition of the hypotheses given in (15a) and (15b) matches the definition in (5) for the special case where the dictionary contains just one signal. In this special case, the signal input \mathbf{f}^* is in the dictionary under the null hypothesis in both DSD and ASD problem formulations.²

5.2 Anomaly detection approach

Our anomaly detection approach and the associated theoretical analysis are based on a “distance preservation” property of \mathbf{A} , which is stated formally in (18). We propose an anomaly detection method that controls the false discovery rate (FDR) below a desired level δ for different background and sensor noise statistics. In other words, we control the expected ratio of falsely declared anomalies to the total number of signals declared to be anomalous. Note that here we work with the FDR as opposed to the pFDR, since it is possible for a scene to not contain any anomalies at all. We let $V/R = 0$ for $R = V = 0$ since one does not declare any signal to be anomalous in this case. In [5], Benjamini and Hochberg discuss a p-value based procedure, “BH procedure”, that controls the false discovery rate of M independent hypothesis tests below a desired level. Let

$$d_i = \min_{\mathbf{f} \in \mathcal{D}} \|\mathbf{y}_i - \alpha_i \mathbf{A} \mathbf{f}\| = \min_{\mathbf{f} \in \mathcal{D}} \|\alpha_i \mathbf{A} (\mathbf{f}_i^* - \mathbf{f}) + \mathbf{n}_i\| \quad (16)$$

be the test statistic at the i^{th} location. The p-value can be defined in terms of our test statistic as follows:

$$p_i = \mathbb{P}(\tilde{d}_i \geq d_i | \mathcal{H}_{0i}) \quad (17)$$

¹Note that τ cannot exceed $\sqrt{2}$ because we assume that all targets of interest, including those in \mathcal{D} and the actual target \mathbf{f}^* , are unit-norm.

²The anomaly detection problem discussed here is more accurately described as target detection in the classical detection theory vocabulary. However, in recent works [23, 46], the authors assume that the nominal distribution is obtained from training data and a test sample is declared to be anomalous if it falls outside of the nominal distribution learned from the training data. Our work is in a similar spirit where we learn our dictionary from training data and label any test spectrum that does not correspond to our dictionary as being anomalous.

where $\tilde{d}_i = \min_{\mathbf{f} \in \mathcal{D}} \|\alpha_i \mathbf{A}(\mathbf{f}_i^* - \mathbf{f}) + \mathbf{n}\|$ and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is independent of \mathbf{n}_i . This is the probability under the null hypothesis, of acquiring a test statistic at least as extreme as the one observed. Let us denote the ordered set of p-values by $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(M)}$ and let $\mathcal{H}_{(0i)}$ be the null hypothesis corresponding to $(i)^{\text{th}}$ p-value. The BH procedure says that if we reject all $\mathcal{H}_{(0i)}$ for $i = 1, \dots, t$ where t is the largest i for which $p_{(i)} \leq i\delta/M$, then the FDR is controlled at δ .

To apply this procedure in our setting, we need to find a tractable expression for the p-value at every location. This can be accomplished when \mathbf{A} satisfies the distance-preservation condition stated below. Let $V = \mathcal{D} \cup \{\mathbf{f}_i^* : i \in \{1, \dots, M\}\}$ be the set of all signals in the dictionary and the ones whose projections are measured. Note that $|V| \leq M + m$. For a given $\epsilon \in (0, 1)$, a projection operator $\mathbf{A} \in \mathbb{R}^{K \times N}$, $K \leq N$, is distance-preserving on V if the following holds for all $u, v \in V$:

$$(1 - \epsilon)\|u - v\| \leq \|\mathbf{A}(u - v)\| \leq (1 + \epsilon)\|u - v\|, \forall u, v \in V. \quad (18)$$

The existence of such projection operators is guaranteed by the celebrated Johnson and Lindenstrauss (JL) lemma [25], which says that there exists random constructions of \mathbf{A} for which (18) holds with probability at least $1 - 2|V|^2 e^{-Kc(\epsilon)}$ provided $K = \mathcal{O}(\log |V|) \leq N$, where $c(\epsilon) = \epsilon^2/16 - \epsilon^3/48$ [1, 4]. Examples of such constructions are: (a) Gaussian matrices whose entries are drawn from $\mathcal{N}(0, 1/K)$, (b) Bernoulli matrices whose entries are $\pm 1/\sqrt{N}$ with probability $1/2$, (c) random matrices whose entries are $\pm \sqrt{3/N}$ with probability $1/6$ and zero with probability $2/3$ [1, 4], and (d) matrices that satisfy the Restricted Isometry Property (RIP) where the signs of the entries in each column are randomized [28].

We now state our main theorem that gives a tight upper bound on the p-value at every location when $\{\alpha_i\}$ are unknown and are estimated from the observations. Let $\{\hat{\alpha}_i\}$ be the estimates of $\{\alpha_i\}$ that satisfy

$$1 - \zeta \leq \frac{\alpha_i}{\hat{\alpha}_i} \leq 1 + \zeta \quad (19)$$

for $i = 1, \dots, M$ where $\zeta \in [0, 1]$ is a measure of the accuracy of the estimation procedure.

Theorem 4. *If the i^{th} hypothesis test is defined according to (15a) and (15b), the projection matrix \mathbf{A} satisfies (18) for a given $\epsilon \in (0, 1)$, and the estimates $\{\hat{\alpha}_i\}$ satisfy (19) for some $\zeta \in [0, 1]$, then the bound*

$$p_i \leq 1 - \mathcal{F}\left(d_i^2; K, (1 + \epsilon)^2 \hat{\alpha}_i^2 (\zeta + \tau)^2\right) \quad (20)$$

holds for all $i = 1, \dots, M$ where $\mathcal{F}(\cdot; K, \nu)$ is the CDF of a noncentral χ^2 random variable with K degrees of freedom and noncentrality parameter ν [54].

The proof of this theorem is given in Appendix D. We find the p-value upper bounds at every location and use the BH procedure to perform anomaly

detection. The performance of this procedure depends on the values of K , $\{\alpha_i\}$, τ and ϵ . The parameter ϵ is a measure of the accuracy with which the projection matrix \mathbf{A} preserves the distances between any two vectors in \mathbb{R}^N . A value of ϵ close to zero implies that the distances are preserved fairly accurately. When $\{\alpha_i\}$ are unknown and estimated from the observations, the performance depends on the accuracy of the estimation procedure, which is reflected in our bounds in (20) through ζ .

One can easily estimate $\{\alpha_i\}$ from $\{\mathbf{y}_i\}$ for some choices of \mathbf{A} . For instance, if the entries of the projection matrix \mathbf{A} are drawn from $\mathcal{N}(0, 1/K)$, the $\{\alpha_i\}$ can be estimated using a maximum likelihood estimator (MLE) by exploiting the statistics of the projection matrix and noise. Note that the j^{th} element of the i^{th} measured spectrum is $y_{i,j} = \sum_{k=1}^N \alpha_i f_{i,k}^* a_{j,k} + n_{i,j} \sim \mathcal{N}\left(0, \sum_{k=1}^N \frac{\alpha_i^2}{K} f_{i,k}^{*2} + 1\right)$ for $j \in \{1, \dots, K\}$. Since $\|\mathbf{f}_i^*\|_2 = 1$ according to our problem formulation, $y_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{\alpha_i^2}{K} + 1\right)$. The MLE of α_i given by $\hat{\alpha}_i = \arg \max_{\alpha} \mathbb{P}(\mathbf{y}_i | \mathbf{A}, \alpha)$ then reduces to

$$\hat{\alpha}_i = \sqrt{(\|\mathbf{y}_i\|^2 - K)}. \quad (21)$$

In practice, we use $\hat{\alpha}_i = \sqrt{(\|\mathbf{y}_i\|^2 - K)_+}$ where the $(a)_+ = a$ if $a \geq 0$ and 0 otherwise to ensure that $\|\mathbf{y}_i\|^2 - K$ is nonnegative. We can use concentration inequalities to show that with high probability, $\|\mathbf{y}_i\|_2^2$ is tightly concentrated around its mean $\mathbb{E}[\|\mathbf{y}_i\|_2^2] = \alpha_i^2 + K$. Since $y_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{\alpha_i^2}{K} + 1\right)$, $\frac{K}{\alpha_i^2 + K} \|\mathbf{y}_i\|_2^2 \sim \chi_K^2$. From Lemma 2.2 in [52], and Proposition 1 and Remark 1 in [51], for any $t > 0$

$$\mathbb{P}\left(\left|\|\mathbf{y}_i\|_2^2 - (\alpha_i^2 + K)\right| \geq t\right) \leq C \exp(-ct^2) \quad (22)$$

for some absolute constants $C, c > 0$. This result shows that with high probability, $\|\mathbf{y}_i\|_2^2 - K$ is nonnegative.

The experimental results discussed in Sec. 6 demonstrate the performance of this detector as a function of K , $\{\alpha_i\}$ and τ when $\{\alpha_i\}$ are known and as a function of K , τ and ζ when $\{\alpha_i\}$ are estimated.

6 Experimental Results

In the experiments that follow, the entries of \mathbf{A} are drawn from $\mathcal{N}(0, 1/K)$.

6.1 Dictionary signal detection

To test the effectiveness of our approach, we formed a dictionary \mathcal{D} of nine spectra (corresponding to different kinds of trees, grass, water bodies and roads) obtained from a labeled HyMap (Hyperspectral Mapper) remote sensing data set [32], and simulated a realistic dataset using the spectra from this dictionary.

Each HyMap spectrum is of length $N = 106$. We generated projection measurements of these data such that $\mathbf{z}_i = \alpha_i \Phi(\mathbf{f}_i^* + \mathbf{b}_i) + \mathbf{w}_i$ according to (1), where $\mathbf{w}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, $\mathbf{f}_i^* \in \mathcal{D}$ for $i = 1, \dots, 8100$, $\mathbf{b}_i \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$ such that $\boldsymbol{\Sigma}_b$ satisfies the condition in (4), and $\alpha_i = \alpha_i^* \sqrt{K}$ where $\alpha_i^* \sim \mathcal{U}[21, 25]$ and \mathcal{U} denotes uniform distribution. We let $\sigma^2 = 5$ and model $\{\alpha_i\}$ to be proportional to \sqrt{K} to account for the fact that the total observed signal energy increases as the number of detectors increases. We transform the \mathbf{z}_i by a series of operations to arrive at a model of the form discussed in (2), which is $\mathbf{y}_i = \alpha_i \mathbf{A} \mathbf{f}_i^* + \mathbf{n}_i$. For this dataset, $p_{\min} = 0.04938$, $p_{\max} = 0.1481$, and $d_{\min} = 0.04341$.

We evaluate the performance of our detector (7) on the transformed observations, relative to the number of measurements K , by comparing the detection results to the ground truth. Our MAP detector returns a label L_i^{MAP} for every observed spectrum which is determined according to

$$L_i^{\text{MAP}} = \arg \min_{\ell \in \{1, \dots, m\}, \mathbf{f}^{(\ell)} \in \mathcal{D}} \left(\frac{1}{2} \|\mathbf{y}_i - \alpha_i \mathbf{A} \mathbf{f}^{(\ell)}\|^2 - \log p^{(\ell)} \right)$$

where m is the number of signals in \mathcal{D} , and $p^{(\ell)}$ is the a priori probability of target class ℓ . In our experiments we evaluate the performance of our classifier when (a) $\{\alpha_i\}$ are known (AK) and (b) $\{\alpha_i\}$ are unknown (AU) and must be estimated from \mathbf{y} , respectively. The empirical pFDR $^{(j)}$ for each target spectrum j is calculated as follows:

$$\text{pFDR}^{(j)} = \frac{\sum_{i=1}^M \mathbb{I}_{\{L_i^{\text{GT}}=j\}} \mathbb{I}_{\{L_i^{\text{MAP}} \neq j\}}}{\sum_{i=1}^M \mathbb{I}_{\{L_i^{\text{MAP}} \neq j\}}}$$

where $\{L_i^{\text{GT}}\}$ denote the ground truth labels. The empirical pFDR $^{(\cdot)}$ is the ratio of the number of missed targets to the total number of signals that were declared to be nontargets. The plots in Fig. 1(a) show the results obtained using our target detection approach under the AK case (shown by a dark gray dashed line) and the AU case (shown by a light gray dashed line), compared to the theoretical upper bound (shown by a solid line). These results are obtained by averaging the pFDR values obtained over 1000 different noise, sensing matrix and background realizations. Note that theoretical results only apply to the AK case since they were derived under the assumption of $\{\alpha_i\}$ being known. The experimental results are shown for both AK and AU cases to provide a comparison between the two scenarios. In both these cases, the worst-case empirical pFDR curves decay with the increase in the values of K . In the AK case, in particular, the worst-case empirical pFDR curve decays at the same rate as the upper bound. In this experiment, for a fixed α_{\min} and d_{\min} , we chose K to satisfy (13c). The theory is somewhat conservative, and in practice the method works well even when the values of K are below the bound in (13c).

In the experiment that follows, we let $\alpha_i^* \sim \mathcal{U}[10, 20]$, where \mathcal{U} denotes a uniform random variable, $\alpha_i = \sqrt{K} \alpha_i^*$ and evaluate the performance of our detector for different values of K that are not necessarily chosen to satisfy

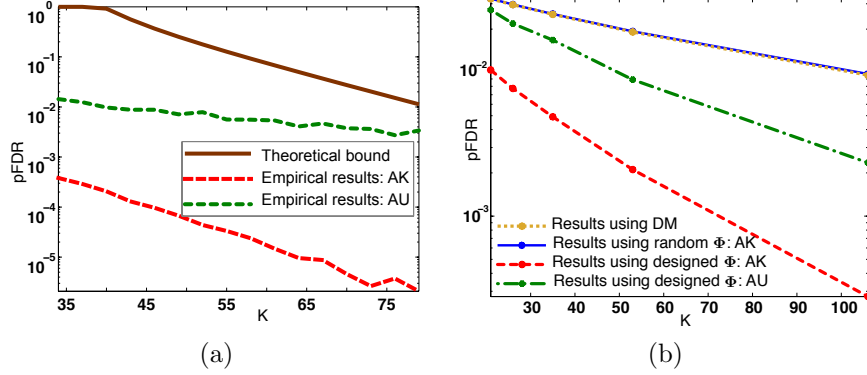


Figure 1: Compressive target detection results under the AK ($\{\alpha_i\}$ known) and AU ($\{\alpha_i\}$ unknown) cases respectively as a function of K . (a) Comparison of the worst-case empirical pFDR curves with the theoretical bounds when SNR is high. (b) Comparison of the results obtained by the proposed method using projection measurements using Φ designed according to (24), Φ chosen at random, and the ones using downsampled measurements (DM) when the SNR is low.

(13c). In addition, we also compare the performance of our detection method to that of a MAP based target detector operating on downsampled versions of our simulated spectral input image. The reason behind such a comparison is to show what kinds of measurements yield better results given a fixed number of detectors.

For an input spectrum $\mathbf{g} \in \mathbb{R}^N$, we let $\tilde{\mathbf{g}} \in \mathbb{R}^K$ denote its downsampled approximation. Specifically, the j^{th} element of $\tilde{\mathbf{g}}$ is $\sum_{\ell=1}^r g_{(j-1)r+\ell}$ where $r = \lceil N/K \rceil$. Let us consider making observations of the form

$$\mathbf{y}_i = \frac{\tilde{\mathbf{g}}_i}{c} + \mathbf{n}_i \in \mathbb{R}^K \quad (23)$$

where $\tilde{\mathbf{g}}_i = \alpha_i \tilde{\mathbf{f}}_i^* + \tilde{\mathbf{b}}_i$ is the K -dimensional downsampled version of $\mathbf{f}_i^* + \mathbf{b}_i$ for $K \leq N$, $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ for $\sigma^2 = 5$ and c is a constant that is chosen to preserve the mean signal-to-noise ratio corresponding to the downsampled and projection measurements. The MAP-based detector operating on the downsampled data returns a label D_i^{MAP} for every observed spectrum which is determined according to

$$D_i^{\text{MAP}} = \arg \min_{\ell \in \{1, \dots, m\}, \mathbf{f}^{(\ell)} \in \mathcal{D}} \left(\mathbf{y}_i - \alpha_i \tilde{\mathbf{f}}^{(\ell)} \right)^T G^{-1} \left(\mathbf{y}_i - \alpha_i \tilde{\mathbf{f}}^{(\ell)} \right) - \log p^{(\ell)}$$

where $G = \tilde{\Sigma}_b + \sigma^2 \mathbf{I}$ and $\tilde{\Sigma}_b$ is the covariance matrix obtained from the downsampled versions of the background training data and $\tilde{\mathbf{f}}^{(\ell)}$ is the downsampled version of $\mathbf{f}^{(\ell)} \in \mathcal{D}$. The algorithm declares that target spectrum $\mathbf{f}^{(j)} \in \mathcal{D}$ is present in the i^{th} location if $D_i^{\text{MAP}} = j$. In order to illustrate the advantages

of using a Φ designed according to (24), we compare the performances of the proposed anomaly detector when Φ is chosen to be a random Gaussian matrix whose entries are drawn from $\mathcal{N}(0, 1/K)$ and when Φ is chosen according to (24). Fig. 1(b) shows a comparison of the results obtained using the projection measurements obtained using Φ designed according to (24), Φ chosen at random, and the downsampled measurements under the AK case. These results show that the detection algorithm operating on projection measurements using Φ designed using background and sensor noise statistics yield significantly better results than the one operating on the downsampled data, and that the empirical pFDR values in our method decays with K . The improvement in performance using projection measurements comes from the distance-preservation property of the projection operator \mathbf{A} . While a Gaussian sensing matrix \mathbf{A} preserves distances between any pair of vectors from a finite collection of vectors with high probability [1, 4], downsampling loses some of the fine differences between similar-looking spectra in the dictionary. Furthermore, when Φ is chosen at random, the resulting whitened transformation matrix is not necessarily distance-preserving. This may worsen the performance as illustrated in Fig. 1(b).

6.2 Anomaly detection

In this section, we evaluate the performance of our anomaly detection method on (a) a simulated dataset and provide a comparison of the results obtained using the proposed projection measurements and the ones obtained using downsampled measurements, and (b) real AVIRIS (Airborne Visible InfraRed Imaging Spectrometer) dataset.

6.2.1 Experiments on simulated data

We simulate a spectral image \mathbf{f}^* composed of 8100 spectra, where each of them is either drawn from a dictionary $\mathcal{D} = \{\mathbf{f}^{(1)}, \dots, \mathbf{f}^{(5)}\}$ consisting of five labeled spectra from the HyMap data that correspond to a natural landscape (trees, grass and lakes) or is anomalous. The anomalous spectrum is extracted from unlabeled AVIRIS data, and the minimum distance between the anomalous spectrum $\mathbf{f}^{(a)}$ and any of the spectra in \mathcal{D} is $d_{\min} = \min_{\mathbf{f} \in \mathcal{D}} \|\mathbf{f} - \mathbf{f}^{(a)}\| = 0.5308$. The simulated data has 625 locations that contain the anomalous spectrum. Our goal is to find the spatial locations that contain the anomalous AVIRIS spectrum given noisy measurements of the form $\mathbf{z}_i = \Phi(\alpha_i \mathbf{f}_i^* + \mathbf{b}_i) + \mathbf{w}_i$ where $\mathbf{b}_i \sim (\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$, Φ is designed according to (24), $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ and $\mathbf{f}_i^* \in \mathcal{D}$ under \mathcal{H}_{0i} . As discussed in Sec. 5, \mathbf{f}_i^* is anomalous under \mathcal{H}_{1i} , and our goal is to control the FDR below a user-specified false discovery level δ . We simulate $\{\alpha_i\} = \sqrt{K} \alpha_i^*$ where $\alpha_i^* \sim \mathcal{U}[2, 3]$. In this experiment we assume the availability of background training data to estimate the background statistics and the sensor noise variance σ^2 . Given the knowledge of the background statistics, we perform the whitening transformation discussed in Sec. 2 and evaluate the detection performance on the preprocessed observations given by (2).

For a fixed $\tau = 0.1$ and $\epsilon = 0.1$, we evaluate the performance of the detector as the number of measurements K increases under the AK and AU cases respectively, by comparing the *pseudo-ROC* (receiver operating characteristic) curves obtained by plotting the empirical false discovery rate against $1 - \text{FNR}$, where FNR is the false nondiscovery rate. Note that $1 - \text{FNR}$ is the expected ratio of the number of null hypotheses that are correctly rejected to the number of declared null hypotheses. The empirical FDR and FNR are computed according to

$$\text{FDR} = \frac{\sum_{i=1}^M \mathbb{I}_{\{L_i^{\text{GT}}=0\}} \mathbb{I}_{\{p_i \leq p_t\}}}{\sum_{i=1}^M \mathbb{I}_{\{p_i \leq p_t\}}} \text{ and } \text{FNR} = \frac{\sum_{i=1}^M \mathbb{I}_{\{L_i^{\text{GT}}=1\}} \mathbb{I}_{\{p_i > p_t\}}}{\sum_{i=1}^M \mathbb{I}_{\{p_i > p_t\}}}$$

where p_t is the p-value threshold such that the BH procedure rejects all null hypotheses for which $p_i \leq p_t$, and the ground truth label $L_i^{\text{GT}} = 0$ if the i^{th} spectrum is not anomalous, and 1 otherwise. In this experiment, we consider three different values of K approximately given by $K \in \{N/6, N/3, N/2\}$ where $N = 106$, and evaluate the performance of our detector for each K . Furthermore, in our experiments with simulated data, we declare a spectrum to be anomalous if $d_i \geq \eta$ where η is a user-specified threshold and d_i is defined in (16). We use the p-value upper bound in (20) in our experiments with real data where the ground truth is unknown.

We compare the performance of our method to a generalized likelihood ratio test (GLRT)-based procedure operating on downsampled data, where we collect measurements of the form in (23) and $\mathbf{f}_i^* \in \mathcal{D}$ under \mathcal{H}_{0i} . Observe that $\mathbf{y}_i | \mathcal{H}_{0i} \sim \sum_{\mathbf{f} \in \mathcal{D}} \mathbb{P}(\mathbf{f}_i^* = \mathbf{f}) \mathcal{N}(\alpha_i \tilde{\mathbf{f}}, \tilde{\Sigma}_b + \mathbf{I})$, where $\tilde{\mathbf{f}}$ refers to the downsampled version of $\mathbf{f} \in \mathcal{D}$. In this experiment we assume that each spectrum in \mathcal{D} is equally likely under \mathcal{H}_{0i} for $i = 1, \dots, M$. The GLRT-based approach declares the i^{th} spectrum to be anomalous if

$$-\log \mathbb{P}(\mathbf{y}_i | \mathcal{H}_{0i}) \underset{\mathcal{H}_{0i}}{\overset{\mathcal{H}_{1i}}{\gtrless}} \eta$$

for $i = 1, \dots, M$, where η is a user-specified threshold [45]. While our anomaly detection method is designed to control the FDR below a user-specified threshold, the GLRT-based method is designed to increase the probability of detection while keeping the probability of false alarm as low as possible. To facilitate a fair evaluation of these methods, we compare the pseudo-ROC curves (FDR versus $1 - \text{FNR}$) and the actual ROC curves (probability of false alarm p_f versus probability of detection p_d) corresponding to these methods obtained by averaging the empirical FDR, FNR, p_d and p_f over 1000 different noise and sensing matrix realizations for different values of K . We also compare the performance of the proposed method when Φ is chosen according to (24) and when it is chosen at random, as discussed in the previous section. Figs. 2(a) and 2(e) show the pseudo-ROC plots and the conventional ROC plots obtained using the GLRT-based method operating on downsampled data when $\{\alpha_i\}$ are known. Figs. 2(b) and 2(f) show the results obtained by using a random Gaussian Φ

instead of the Φ in (24). Figs. 2(c) and 2(g) show the pseudo-ROC plots and the conventional ROC plots obtained using our method when $\{\alpha_i\}$ are known. These plots show that performing anomaly detection from our *designed* projection measurements yields better results than performing anomaly detection on downsampled measurements and on measurements obtained using a random Gaussian Φ . This is largely due to the fact that carefully chosen projection measurements preserve distances (up to a constant factor) among pairs of vectors in a finite collection, where as the downsampled measurements fail to preserve distances among vectors that are very similar to each other. Similarly, a random projection matrix Φ is not necessarily distance-preserving post-whitening transformation, which leads to poor performance as illustrated in Figs. 2(b) and 2(f). Figs. 2(d) and 2(h) show the pseudo-ROC plots and the conventional ROC plots obtained using our method when $\{\alpha_i\}$ are unknown, and are estimated from the measurements. Note that the value of ζ decreases as K increases since the estimation accuracy of $\{\alpha_i\}$ increases with increase in K . These plots show that the performance improves as we collect more observations, and that, as expected, the performance under the AK case is better than the performance under the AU case.

6.2.2 Experiments on real AVIRIS data

To test the performance of our anomaly detector on a real dataset, we consider the unlabeled AVIRIS Jasper Ridge dataset $\mathbf{g} \in \mathbb{R}^{614 \times 512 \times 197}$, which is publicly available from the NASA AVIRIS website, <http://aviris.jpl.nasa.gov/html/aviris.freedata.html>. We split this data spatially to form equi-sized training and validation datasets, \mathbf{g}^t and \mathbf{g}^v respectively, each of which is of size $128 \times 128 \times 197$. Figs. 3(a) and 3(b) show images of the AVIRIS training and validation data summed through the spectral coordinates. The training data are comprised of a rocky terrain with a small patch of trees. The validation data seems to be made of a similar rocky terrain, but also contain an anomalous lake-like structure. The goal is to evaluate the performance of the detector in detecting the anomalous region in the validation data for different values of K . We cluster the spectral targets in the normalized training data to eight different clusters using the K-means clustering algorithm and form a dictionary \mathcal{D} comprising of the cluster centroids. Given the dictionary and the validation data, we find the ground truth by labeling the i^{th} validation spectrum as anomalous if $\min_{\mathbf{f} \in \mathcal{D}} \left\| \mathbf{f} - \frac{\mathbf{g}_i^v}{\|\mathbf{g}_i^v\|} \right\| > \tau$. Since the statistics of the possible background contamination in the data could not be learned in this experiment because of the lack of labeled training data, the dictionary might be background contaminated as well. The parameter τ encapsulates this uncertainty in our knowledge of the dictionary. In this experiment, we set $\tau = 0.2$.

We generate measurements of the form $\mathbf{y}_i = \sqrt{K} \mathbf{g}_i^v + \mathbf{n}_i$ for $i = 1, \dots, 128 \times 128$, where $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The \sqrt{K} factor indicates that the observed signal strength increases with K . For a fixed FDR control value of 0.01, Figs. 3(c) and 3(d) show the results obtained for $K \approx N/5$ and $K \approx N/2$ respectively.

Fig. 3(e) shows how the probability of error decays as a function of the number of measurements K . The results presented here are obtained by averaging over 1000 different noise and sensing matrix realizations. From these results, we can see that the number of detected anomalies increases with K and the number of misclassifications decrease with K .

7 Conclusion

This work presents computationally efficient approaches for detecting known targets and anomalies of different strengths from projection measurements without performing a complete reconstruction of the underlying signals, and offers theoretical bounds on the worst-case target detector performance. This paper treats each signal as independent of its spatial or temporal neighbors. This assumption is reasonable in many contexts, especially when the spatial or temporal resolution is low relative to the spatial homogeneity of the environment or the pace with which a scene changes. However, emerging technologies in computational optical systems continue to improve the resolution of spectral imagers. In our future work we will build upon the methods that we have discussed here to exploit the spatial or temporal correlations in the data.

A Proof of Theorem 1

Using linear algebra and matrix theory, it is possible to show that if $\mathbf{B} = \mathbf{I} - \mathbf{A}\mathbf{\Sigma}_b\mathbf{A}^T$ is positive definite, then

$$\mathbf{\Phi} = \sigma\mathbf{B}^{-1/2}\mathbf{A} \quad (24)$$

satisfies (3).³ In particular, we can substitute (24) in (3) to verify that the proposed construction of $\mathbf{\Phi}$ satisfies (3). Observe that $\mathbf{C}_{\mathbf{\Phi}} = (\mathbf{\Phi}\mathbf{\Sigma}_b\mathbf{\Phi}^T + \sigma^2\mathbf{I})^{-1/2}$ can be written in terms of (24) as follows:

$$\begin{aligned} \mathbf{C}_{\mathbf{\Phi}} &= \left(\left[\sigma\mathbf{B}^{-\frac{1}{2}}\mathbf{A} \right] \mathbf{\Sigma}_b \left[\sigma\mathbf{B}^{-\frac{1}{2}}\mathbf{A} \right]^T + \sigma^2\mathbf{I} \right)^{-\frac{1}{2}} \\ &= \left(\sigma^2\mathbf{B}^{-1/2} (\mathbf{A}\mathbf{\Sigma}_b\mathbf{A}^T) \left(\mathbf{B}^{-\frac{1}{2}} \right)^T + \sigma^2\mathbf{I} \right)^{-\frac{1}{2}} \\ &= \left(\sigma^2\mathbf{B}^{-\frac{1}{2}} (\mathbf{I} - \mathbf{B}) \left(\mathbf{B}^{-\frac{1}{2}} \right)^T + \sigma^2\mathbf{I} \right)^{-\frac{1}{2}} = (\sigma^2\mathbf{B}^{-1})^{-\frac{1}{2}} = \sigma^{-1}\mathbf{B}^{\frac{1}{2}} \quad (25) \end{aligned}$$

where the third-to-last equation follows from the definition of \mathbf{B} and (25) follows from the fact that \mathbf{B} is symmetric and positive definite. If \mathbf{B} is positive definite, then \mathbf{B}^{-1} is positive definite as well and can be decomposed

³The authors would like to thank Prof. Roummel Marcia for fruitful discussions related to this point.

as $\mathbf{B}^{-1} = (\mathbf{B}^{-1/2})^T \mathbf{B}^{-1/2}$, where the matrix square root $\mathbf{B}^{-1/2}$ is symmetric and positive definite. By substituting (25) and (24) in (3), we have $\mathbf{C}_\Phi \Phi = \sigma^{-1} \mathbf{B}^{1/2} \sigma \mathbf{B}^{-1/2} \mathbf{A} = \mathbf{A}$. A sufficient condition for \mathbf{B} to be positive definite can be derived as follows.

To ensure positive definiteness of \mathbf{B} , we must have

$$\mathbf{x}^T \mathbf{B} \mathbf{x} = \mathbf{x}^T \mathbf{x} - \mathbf{x}^T (\mathbf{A} \Sigma_{\mathbf{b}} \mathbf{A}^T) \mathbf{x} > 0 \quad (26)$$

for any nonzero $\mathbf{x} \in \mathbb{R}^K$. Note that since $\Sigma_{\mathbf{b}}$ is positive semidefinite, $\mathbf{x}^T (\mathbf{A} \Sigma_{\mathbf{b}} \mathbf{A}^T) \mathbf{x} \geq 0$. However, the right hand side of (26) is > 0 only if the spectral norm of $\mathbf{A} \Sigma_{\mathbf{b}} \mathbf{A}^T$ is < 1 , since $\mathbf{x}^T (\mathbf{A} \Sigma_{\mathbf{b}} \mathbf{A}^T) \mathbf{x} \leq \|\mathbf{x}\|^2 \cdot \|\mathbf{A} \Sigma_{\mathbf{b}} \mathbf{A}^T\|$. The norm of $\mathbf{A} \Sigma_{\mathbf{b}} \mathbf{A}^T$ is in turn bounded above by

$$\|\mathbf{A} \Sigma_{\mathbf{b}} \mathbf{A}^T\| \leq \|\mathbf{A}\| \|\Sigma_{\mathbf{b}}\| \|\mathbf{A}^T\| = \|\mathbf{A}\|^2 \|\Sigma_{\mathbf{b}}\| = \|\mathbf{A}\|^2 \lambda_{\max}$$

since $\|\mathbf{A}\| = \|\mathbf{A}^T\|$ and $\|\Sigma_{\mathbf{b}}\| = \lambda_{\max}$, where λ_{\max} is the largest eigenvalue of $\Sigma_{\mathbf{b}}$. To ensure $\|\mathbf{A} \Sigma_{\mathbf{b}} \mathbf{A}^T\| < 1$, $\|\mathbf{A}\|^2 \lambda_{\max}$ has to be < 1 , which leads to the result of Theorem 1.

B Proof of Theorem 2

The proof of Theorem 2 adapts the proof techniques from [48] to nonidentical independent hypothesis tests. We begin by expanding the pFDR definition in (8) as follows:

$$\text{pFDR}^{(j)}(\Gamma) = \sum_{k=1}^M \mathbb{E} \left[\frac{V(\Gamma)}{R(\Gamma)} \middle| R(\Gamma) = k \right] \times \mathbb{P}(R(\Gamma) = k | R(\Gamma) > 0).$$

Observe that $R(\Gamma) = k$ implies that there exists some subset $S_k = \{u_1, \dots, u_k\} \subseteq \{1, \dots, M\}$ of size k such that $\mathbf{y}_{u_\ell} \in \Gamma_{u_\ell}^{(j)}$ for $\ell = 1, \dots, k$ and $\mathbf{y}_i \notin \Gamma_i^{(j)}$ for all $i \notin S_k$. To simplify the notation, let $\Lambda_{S_k} = \prod_{u \in S_k} \Gamma_u^j \times \prod_{\ell \notin S_k} \tilde{\Gamma}_\ell^{(j)}$, where $\tilde{\Gamma}_\ell^{(j)}$ is the complement of $\Gamma_\ell^{(j)}$, denote the significance region that corresponds to set S_k , and $\mathbf{T} = (\mathbf{y}_1, \dots, \mathbf{y}_M)$ be a set of test statistics corresponding to each hypothesis test. Considering all such subsets we have

$$\text{pFDR}^{(j)}(\Gamma) = \sum_{k=1}^M \sum_{S_k} \mathbb{E} \left[\frac{V(\Gamma)}{k} \middle| \mathbf{T} \in \Lambda_{S_k} \right] \times \mathbb{P}(\mathbf{T} \in \Lambda_{S_k} | R(\Gamma) > 0). \quad (27)$$

By plugging in the definition of $V(\{\Gamma_i\})$ from (9), we have

$$\begin{aligned} \mathbb{E}[V(\Gamma) | \mathbf{T} \in \Lambda_{S_k}] &= \mathbb{E} \left[\sum_{i=1}^M \mathbb{I}_{\{\mathbf{y}_i \in \Gamma_i^{(j)}\}} \mathbb{I}_{\{\mathcal{H}_i^{(j)} = 0\}} \middle| \mathbf{T} \in \Lambda_{S_k} \right] \\ &= \sum_{\ell=1}^k \mathbb{E} \left[\mathbb{I}_{\{\mathcal{H}_{u_\ell}^{(j)} = 0\}} \middle| \mathbf{y}_{u_\ell} \right] = \sum_{\ell=1}^k \mathbb{P}(\mathcal{H}_{u_\ell}^{(j)} = 0 | \mathbf{y}_{u_\ell} \in \Gamma_{u_\ell}^{(j)}) \end{aligned} \quad (28)$$

for all $u_\ell \in S_k$ since the tests are independent of each other given \mathbf{A} . The posterior probability $\mathbb{P}\left(\mathcal{H}_i^{(j)} = 0 \mid \mathbf{y}_i \in \Gamma_i^{(j)}\right)$ for the i^{th} hypothesis test can be expanded using Bayes' rule as

$$\begin{aligned} \mathbb{P}\left(\mathcal{H}_{0i}^{(j)} \mid \mathbf{y}_i \in \Gamma_i^{(j)}\right) &= \frac{\mathbb{P}\left(\mathbf{y}_i \in \Gamma_i^{(j)} \mid \mathcal{H}_{0i}\right) \mathbb{P}\left(\mathcal{H}_{0i}^{(j)}\right)}{\mathbb{P}\left(\mathbf{y}_i \in \Gamma_i^{(j)}\right)} \\ &\equiv \frac{\mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}^{(j)} \mid \mathbf{f}_i^* = \mathbf{f}^{(j)}\right) \mathbb{P}\left(\mathbf{f}_i^* = \mathbf{f}^{(j)}\right)}{\mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}^{(j)}\right)}, \end{aligned} \quad (29)$$

where $\hat{\mathbf{f}}_i = \arg \max_{\mathbf{f}^{(\ell)} \in \mathcal{D}} \mathbb{P}\left(\mathbf{f}_i^* = \mathbf{f}^{(\ell)} \mid \mathbf{y}_i, \alpha_i, \mathbf{A}\right)$. To upper bound the numerator of (29), consider the probability of misclassification given by $(\mathbf{P}_e)_i = \mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}_i^*\right)$ where $\mathbf{f}_i^* = \mathbf{f}^{(j)} \in \mathcal{D}$, which can be expanded as follows:

$$\begin{aligned} (\mathbf{P}_e)_i &= \mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}_i^*\right) = \sum_{\ell=1}^m \mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}_i^* \mid \mathbf{f}_i^* = \mathbf{f}^{(\ell)}\right) \mathbb{P}\left(\mathbf{f}_i^* = \mathbf{f}^{(\ell)}\right) \\ &\equiv \sum_{\ell=1}^m \mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}^{(\ell)} \mid \mathbf{f}_i^* = \mathbf{f}^{(\ell)}\right) \mathbb{P}\left(\mathbf{f}_i^* = \mathbf{f}^{(\ell)}\right) \geq \mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}^{(j)} \mid \mathbf{f}_i^* = \mathbf{f}^{(j)}\right) \mathbb{P}\left(\mathbf{f}_i^* = \mathbf{f}^{(j)}\right). \end{aligned} \quad (30)$$

The denominator term in (29) can be expanded as follows:

$$\begin{aligned} \mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}^{(j)}\right) &= \mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}^{(j)} \mid \mathbf{f}_i^* = \mathbf{f}^{(j)}\right) \mathbb{P}\left(\mathbf{f}_i^* = \mathbf{f}^{(j)}\right) \\ &\quad + \mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}^{(j)} \mid \mathbf{f}_i^* \neq \mathbf{f}^{(j)}\right) \mathbb{P}\left(\mathbf{f}_i^* \neq \mathbf{f}^{(j)}\right). \end{aligned}$$

Observe that $\mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}^{(j)} \mid \mathbf{f}_i^* = \mathbf{f}^{(j)}\right)$ is nonnegative, and

$$\begin{aligned} \mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}^{(j)} \mid \mathbf{f}_i^* \neq \mathbf{f}^{(j)}\right) &= \mathbb{P}\left(\hat{\mathbf{f}}_i \in \mathcal{D} \setminus \mathbf{f}^{(j)} \mid \mathbf{f}_i^* \neq \mathbf{f}^{(j)}\right) \geq \mathbb{P}\left(\hat{\mathbf{f}}_i = \mathbf{f}_i^* \mid \mathbf{f}_i^* \neq \mathbf{f}^{(j)}\right) \\ &= 1 - \mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}_i^* \mid \mathbf{f}_i^* \neq \mathbf{f}^{(j)}\right) = 1 - \frac{\mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}_i^*, \mathbf{f}_i^* \neq \mathbf{f}^{(j)}\right)}{\mathbb{P}\left(\mathbf{f}_i^* \neq \mathbf{f}^{(j)}\right)} \geq 1 - \frac{\mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}_i^*\right)}{\mathbb{P}\left(\mathbf{f}_i^* \neq \mathbf{f}^{(j)}\right)} \\ &= 1 - \frac{(\mathbf{P}_e)_i}{1 - p^{(j)}}. \end{aligned}$$

Thus

$$\mathbb{P}\left(\hat{\mathbf{f}}_i \neq \mathbf{f}^{(j)}\right) \geq \left(1 - \frac{(\mathbf{P}_e)_i}{1 - p^{(j)}}\right) \left(1 - p^{(j)}\right) = 1 - p^{(j)} - (\mathbf{P}_e)_i. \quad (31)$$

Substituting (30) and (31) in (29),

$$\mathbb{P}\left(\mathcal{H}_{0i}^{(j)} \mid \mathbf{y}_i \in \Gamma_i^{(j)}\right) \leq \frac{(\mathbf{P}_e)_i}{1 - p^{(j)} - (\mathbf{P}_e)_i} \leq \frac{(\mathbf{P}_e)_{\max}}{1 - p^{(j)} - (\mathbf{P}_e)_{\max}}. \quad (32)$$

By substituting (32) in (27) and (28) we have:

$$\begin{aligned} \text{pFDR}^{(j)}(\mathbf{\Gamma}) &\leq \sum_{k=1}^M \sum_{S_k} \frac{1}{k} \left(\sum_{\ell=1}^k \frac{(\mathbf{P}_e)_{\max}}{1 - p^{(j)} - (\mathbf{P}_e)_{\max}} \right) \times \mathbb{P}(\mathbf{T} \in \Lambda_{S_k} | R(\mathbf{\Gamma}) > 0) \\ &= \frac{(\mathbf{P}_e)_{\max}}{1 - p^{(j)} - (\mathbf{P}_e)_{\max}} \sum_{k=1}^M \sum_{S_k} \mathbb{P}(\mathbf{T} \in \Lambda_{S_k} | R(\mathbf{\Gamma}) > 0) \leq \frac{(\mathbf{P}_e)_{\max}}{1 - p^{(j)} - (\mathbf{P}_e)_{\max}} \end{aligned}$$

since $\sum_{k=1}^M \sum_{S_k} \mathbb{P}(\mathbf{T} \in \Lambda_{S_k} | R(\mathbf{\Gamma}) > 0) \leq 1$. The result of Theorem 2 is obtained by finding an upper bound on the worst-case pFDR given by

$$\begin{aligned} \text{pFDR}_{\max} &= \max_{j \in \{1, \dots, m\}} \text{pFDR}^{(j)}(\mathbf{\Gamma}) \\ &\leq \max_{j \in \{1, \dots, m\}} \frac{(\mathbf{P}_e)_{\max}}{1 - p^{(j)} - (\mathbf{P}_e)_{\max}} = \frac{(\mathbf{P}_e)_{\max}}{1 - p_{\max} - (\mathbf{P}_e)_{\max}} \end{aligned}$$

where $p_{\max} = \max_{\ell \in \{1, \dots, m\}} p^{(\ell)}$.

C Proof of Theorem 3

The proof is via a random selection technique, similar to random coding arguments common in information theory. Specifically, we will draw a $K \times N$ sensing matrix \mathbf{A} at random from a particular distribution and then show that, for ϵ , N , and K satisfying the conditions of the theorem, the probability that the conclusions of the theorem will fail to hold for this randomly chosen \mathbf{A} is strictly smaller than unity. This will imply that the conclusions of the theorem must be true for at least one (deterministic) realization of \mathbf{A} .

We begin by specifying all the relevant random variables:

- $\mathbf{f}_1^*, \dots, \mathbf{f}_M^*$ are i.i.d. random variables taking values in the dictionary $\mathcal{D} = \{\mathbf{f}^{(1)}, \dots, \mathbf{f}^{(m)}\}$ with probabilities $p^{(j)} = \Pr\{\mathbf{f}_i^* = \mathbf{f}^{(j)}\}, j \in \{1, \dots, m\}$;
- $\mathbf{n}_1, \dots, \mathbf{n}_M \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$;
- \mathbf{G} is a random $K \times N$ matrix with i.i.d. $\mathcal{N}(0, 1)$ entries.

We assume that $\{\mathbf{f}_i^*\}_{i=1}^M$, $\{\mathbf{n}_i\}_{i=1}^M$, and \mathbf{G} are mutually independent, and we will denote by \mathbf{P} their joint probability distribution. Finally, we let $\mathbf{A} = \frac{1}{\sqrt{K}} \mathbf{G}$ and consider the observation model

$$\mathbf{y}_i = \alpha_i \mathbf{A} \mathbf{f}_i^* + \mathbf{n}_i, \quad i \in \{1, \dots, M\} \quad (33)$$

where $\alpha_1, \dots, \alpha_M > 0$ are the given signal strengths.

We first consider the case when $\alpha_1 = \dots = \alpha_M = \alpha$. Given ϵ , N , and K , we define the following two error events:

$$\mathcal{E}_1 \triangleq \left\{ \|\mathbf{G}\| \geq (1 + \epsilon)(\sqrt{K} + \sqrt{N}) \right\}, \text{ and } \mathcal{E}_2 \triangleq \left\{ \hat{\mathbf{f}}_1 \neq \mathbf{f}_1^* \right\},$$

where, for each $i \in \{1, \dots, M\}$, $\hat{\mathbf{f}}_i$ is defined according to (12). Note that, since we have assumed that the α_i 's are equal and all the pairs $(\mathbf{f}_i^*, \mathbf{n}_i), i \in \{1, \dots, M\}$, are i.i.d.,

$$\mathbb{P}(\hat{\mathbf{f}}_i \neq \mathbf{f}_i^* | \mathbf{A}) = \mathbb{P}(\mathcal{E}_2 | \mathbf{A}), \quad \forall i \in \{1, \dots, M\}. \quad (34)$$

We will now prove that

$$\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) \leq \frac{1 - p_{\min}}{p_{\min}} \left(1 + \frac{\alpha^2 d_{\min}^2}{4K\sigma^2}\right)^{-\frac{K}{2}} + 2 \exp\left(-\frac{(K+N)\epsilon^2}{2}\right). \quad (35)$$

The union bound gives $\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) \leq \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2)$. First, we bound $\mathbb{P}(\mathcal{E}_1)$. To do that, we use the following concentration result for Gaussian random matrices [13]: for any $t \geq 0$,

$$\Pr\left\{\|\mathbf{G}\| \geq \sqrt{K} + \sqrt{N} + t\right\} \leq 2e^{-t^2/2}.$$

Letting $t = \epsilon(\sqrt{K} + \sqrt{N})$ and using the fact that $t^2 \geq (K+N)\epsilon^2$, we get

$$\mathbb{P}(\mathcal{E}_1) \leq 2 \exp\left(-\frac{(K+N)\epsilon^2}{2}\right). \quad (36)$$

Next, we bound $\mathbb{P}(\mathcal{E}_2)$. To that end, we use the following result, which is a straightforward extension of Theorem 1 in [21] to nonequiprobable dictionary elements:

Lemma 1 (Compressive classification error). *Consider the problem of classifying a signal of interest $\mathbf{f}^* \in \mathcal{D} = \{\mathbf{f}^{(1)}, \dots, \mathbf{f}^{(m)}\}$ to one of m known target classes by making observations of the form $\mathbf{y} = \alpha \mathbf{A} \mathbf{f}^* + \mathbf{n}$ where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, given the knowledge of the dictionary \mathcal{D} , prior probabilities $p^{(j)}$ for $j \in \{1, \dots, m\}$, sensing matrix \mathbf{A} , and the noise variance σ^2 . If the entries of \mathbf{A} are drawn i.i.d. from $\mathcal{N}(0, 1/K)$ independently of \mathbf{f}^* and \mathbf{n} , and the estimate $\hat{\mathbf{f}}$ is obtained according to (12), then*

$$\mathbb{P}(\hat{\mathbf{f}} \neq \mathbf{f}^*) \leq \frac{1 - p_{\min}}{p_{\min}} \left(1 + \frac{\alpha^2 d_{\min}^2}{4K\sigma^2}\right)^{-\frac{K}{2}}$$

where the probability is taken with respect to the distributions underlying \mathbf{f}^* , \mathbf{A} , and \mathbf{n} .

Using the above lemma, we have

$$\mathbb{P}(\mathcal{E}_2) \leq \frac{1 - p_{\min}}{p_{\min}} \left(1 + \frac{\alpha^2 d_{\min}^2}{4K}\right)^{-\frac{K}{2}}. \quad (37)$$

Combining (36) and (37), we get (35).

Because of (13a), the right-hand side of (35) is less than $1 - \epsilon - p_{\max}$, which is strictly positive by hypothesis. Thus, from the fact that

$$\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) = \mathbb{E}[\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2 | \mathbf{A})]$$

and from (34), it follows that there exists at least one deterministic choice of the $K \times N$ sensing matrix \mathbf{A}^* , such that:

$$\|\mathbf{A}^*\| \leq (1 + \epsilon) \left(1 + \sqrt{\frac{N}{K}} \right) \quad (38a)$$

$$(\mathbb{P}_e)_{\max}(\mathbf{A}^*) \leq \frac{1 - p_{\min}}{p_{\min}} \left(1 + \frac{\alpha^2 d_{\min}^2}{4K} \right)^{-\frac{K}{2}} + 2 \exp \left(-\frac{(K + N)\epsilon^2}{2} \right) \quad (38b)$$

where, for a given choice of \mathbf{A} , $(\mathbb{P}_e)_{\max}(\mathbf{A})$ denotes the maximum probability of error defined in Theorem 2.

Next, from (38a) and (13b) it follows that \mathbf{A}^* satisfies the conditions of Theorem 1. Finally, we use (11) to bound the worst-case pFDR achievable with \mathbf{A}^* . First of all, we note that the function $U(x) = \frac{x}{1 - p_{\max} - x}$ is twice differentiable and convex on the interval $[0, 1 - p_{\max}]$. Therefore, for any $x \in [0, 1 - p_{\max}]$ and any $h > 0$ small enough so that $x + h \in [0, 1 - p_{\max}]$, we have

$$U(x + h) \leq U(x) + U'(x)h = U(x) + \frac{(1 - p_{\max})h}{(1 - p_{\max} - x - h)^2}. \quad (39)$$

Let us choose

$$x = \frac{1 - p_{\min}}{p_{\min}} \left(1 + \frac{\alpha^2 d_{\min}^2}{4K} \right)^{-\frac{K}{2}} \quad \text{and} \quad h = 2 \exp \left(-\frac{(K + N)\epsilon^2}{2} \right).$$

Then from (13a) we have $x + h \leq 1 - \epsilon - p_{\max} < 1 - p_{\max}$, and from (13c) we have $x + h \geq 0$. Hence, using (39) and simplifying, we obtain the bound

$$\begin{aligned} \text{pFDR}_{\max}(\mathbf{A}^*) &\leq \frac{1}{p_{\min}} \left(\frac{1 - p_{\max}}{1 - p_{\min}} \left(1 + \frac{\alpha^2 d_{\min}^2}{4K} \right)^{\frac{K}{2}} - \frac{1}{p_{\min}} \right)^{-1} + \\ &\quad \frac{2(1 - p_{\max})}{\epsilon^2} \exp \left(-\frac{(K + N)\epsilon^2}{2} \right). \end{aligned}$$

This proves the theorem for the case $\alpha_1 = \dots = \alpha_M = \alpha$.

To handle the case when the α_i 's are distinct, we simply let

$$i^* \triangleq \arg \min_{i \in \{1, \dots, M\}} \alpha_i$$

and replace the definition of the error event \mathcal{E}_2 with $\mathcal{E}'_2 = \{\hat{\mathbf{f}}_{i^*} \neq \mathbf{f}_{i^*}^*\}$. Then the same argument goes through, except that instead of (34) we use the bound

$$\mathbb{P}(\hat{\mathbf{f}}_i \neq \mathbf{f}_i^* | \mathbf{A}) \leq \mathbb{P}(\hat{\mathbf{f}}_{i^*} \neq \mathbf{f}_{i^*}^* | \mathbf{A}) = \mathbb{P}(\mathcal{E}'_2 | \mathbf{A}), \quad \forall i \neq i^*$$

which follows from the following argument. First of all, we can replace the observation model with the equivalent model

$$\tilde{\mathbf{y}}_i = \mathbf{A}\mathbf{f}_i^* + \tilde{\mathbf{n}}_i, \quad i \in \{1, \dots, M\}$$

where $\tilde{\mathbf{n}}_i = \frac{1}{\alpha_i}\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{\alpha_i^2}\mathbf{I})$. Secondly, from the fact that $\alpha_i \geq \alpha_{i^*} \equiv \alpha_{\min}$ for any $i \neq i^*$ it follows that $\tilde{\mathbf{n}}_{i^*}$ is equal in distribution to $\tilde{\mathbf{n}}_i + \tilde{\mathbf{n}}'_i$, where $\tilde{\mathbf{n}}'_i \sim \mathcal{N}(\mathbf{0}, (\frac{1}{\alpha_i^2} - \frac{1}{\alpha_{\min}^2})\mathbf{I})$ is independent of $\tilde{\mathbf{n}}_i$. This implies that the i^* th observation is the noisiest, and the corresponding MAP estimate $\hat{\mathbf{f}}_{i^*}$ has the largest probability of error.

D Proof of Theorem 4

We first prove this theorem assuming that $\{\alpha_i\}$ are known and later extend to the case where $\{\hat{\alpha}_i\}$ are estimated from the observations. Let $\tilde{\mathbf{f}}_i = \arg \min_{\mathbf{f} \in \mathcal{D}} \|\mathbf{f}_i^* - \mathbf{f}\|$. The p-value expression in (17) can be expanded as follows:

$$\begin{aligned} p_i &= \mathbb{P}(\tilde{d}_i \geq d_i \mid \mathcal{H}_{0i}) = \mathbb{P}\left(\min_{\mathbf{f} \in \mathcal{D}} \|\alpha_i \mathbf{A}(\mathbf{f}_i^* - \mathbf{f}) + \mathbf{n}\| \geq d_i \mid \mathcal{H}_{0i}\right) \\ &\leq \mathbb{P}\left(\|\alpha_i \mathbf{A}(\mathbf{f}_i^* - \tilde{\mathbf{f}}_i) + \mathbf{n}\| \geq d_i \mid \mathcal{H}_{0i}\right) = \mathbb{P}\left(\|\alpha_i \mathbf{A}(\mathbf{f}_i^* - \tilde{\mathbf{f}}_i) + \mathbf{n}\|^2 \geq d_i^2 \mid \mathcal{H}_{0i}\right). \end{aligned} \quad (40)$$

Note that $\|\alpha_i \mathbf{A}(\mathbf{f}_i^* - \tilde{\mathbf{f}}_i) + \mathbf{n}\|^2$ is a noncentral χ^2 random variable with K degrees of freedom and a noncentrality parameter $\nu_i = \|\alpha_i \mathbf{A}(\mathbf{f}_i^* - \tilde{\mathbf{f}}_i)\|^2$. Thus (40) can be written in terms of a noncentral χ^2 CDF $\mathcal{F}(d_i^2; K, \nu_i)$ with parameter d_i^2 . The upper and lower bounds on ν_i can be obtained using the properties of the projection matrix \mathbf{A} . Applying (18), we see that

$$\alpha_i^2(1 - \epsilon)^2 \|\mathbf{f}_i^* - \tilde{\mathbf{f}}_i\|^2 \leq \nu_i \leq \alpha_i^2(1 + \epsilon)^2 \|\mathbf{f}_i^* - \tilde{\mathbf{f}}_i\|^2$$

with high probability. Thus,

$$\begin{aligned} p_i &\leq 1 - \mathbb{P}\left(\|\alpha_i \mathbf{A}(\mathbf{f}_i^* - \tilde{\mathbf{f}}_i) + \mathbf{n}\|^2 \leq d_i^2 \mid \mathcal{H}_{0i}\right) = 1 - \mathcal{F}(d_i^2; K, \nu_i) \\ &\leq 1 - \mathcal{F}\left(d_i^2; K, \alpha_i^2(1 + \epsilon)^2 \|\mathbf{f}_i^* - \tilde{\mathbf{f}}_i\|^2\right) \leq 1 - \mathcal{F}\left(d_i^2; K, \alpha_i^2(1 + \epsilon)^2 \tau^2\right) \end{aligned} \quad (41)$$

since $\|\mathbf{f}_i^* - \mathbf{f}\| \leq \tau$ for all $\mathbf{f} \in \mathcal{D}$ under \mathcal{H}_{0i} .

When $\{\alpha_i\}$ are estimated from the observations such that $\{\hat{\alpha}_i\}$ satisfy (19), we can write the p-value expression in (41) as follows:

$$\begin{aligned} p_i &\leq 1 - \mathcal{F}\left(d_i^2; K, \left\|\mathbf{A}\left(\alpha_i \mathbf{f}_i^* - \hat{\alpha}_i \tilde{\mathbf{f}}_i\right)\right\|^2\right) \\ &\leq 1 - \mathcal{F}\left(d_i^2; K, (1 + \epsilon)^2 \hat{\alpha}_i^2 \left\|\frac{\alpha_i}{\hat{\alpha}_i} \mathbf{f}_i^* - \tilde{\mathbf{f}}_i\right\|^2\right) \end{aligned} \quad (42)$$

where (42) is due to the distance preservation property of \mathbf{A} given in (18). Observe that $\left\| \frac{\alpha_i}{\hat{\alpha}_i} \mathbf{f}_i^* - \tilde{\mathbf{f}}_i \right\|^2$ can be upper bounded as shown below:

$$\begin{aligned} \left\| \frac{\alpha_i}{\hat{\alpha}_i} \mathbf{f}_i^* - \tilde{\mathbf{f}}_i \right\|^2 &= \left\| \left(\frac{\alpha_i}{\hat{\alpha}_i} - 1 \right) \mathbf{f}_i^* + \mathbf{f}_i^* - \tilde{\mathbf{f}}_i \right\|^2 \leq \left(\left\| \left(\frac{\alpha_i}{\hat{\alpha}_i} - 1 \right) \mathbf{f}_i^* \right\| + \|\mathbf{f}_i^* - \tilde{\mathbf{f}}_i\| \right)^2 \\ &= \left(\left| \frac{\alpha_i}{\hat{\alpha}_i} - 1 \right| + \|\mathbf{f}_i^* - \tilde{\mathbf{f}}_i\| \right)^2 \leq (\zeta + \|\mathbf{f}_i^* - \tilde{\mathbf{f}}_i\|)^2 \end{aligned}$$

where third-to-last equation is due to the triangle inequality, second-to-last equation comes from the assumption that $\|\mathbf{f}_i^*\| = 1$, and the last inequality is due to (19). By applying this result to (42) and exploiting the fact that $\|\mathbf{f}_i^* - \mathbf{f}\| \leq \tau$ under \mathcal{H}_{0i} for some $\mathbf{f} \in \mathcal{D}$, we have

$$p_i \leq 1 - \mathcal{F} \left(d_i^2; K, (1 + \epsilon)^2 \hat{\alpha}_i^2 \left(\zeta + \|\mathbf{f}_i^* - \tilde{\mathbf{f}}_i\| \right)^2 \right) \leq 1 - \mathcal{F} \left(d_i^2; K, (1 + \epsilon)^2 \hat{\alpha}_i^2 (\zeta + \tau)^2 \right).$$

References

- [1] D. Achlioptas. Database-friendly random projections. In *Proc. 20th ACM Symp. Principles of Database Systems*, pages 274–281. ACM Press, 2001.
- [2] S. Aeron, V. Saligrama, and M. Zhao. Information theoretic bounds for compressed sensing. *Information Theory, IEEE Transactions on*, 56(10):5111–5130, 2010.
- [3] E. Arias-Castro and Y. Eldar. Noise folding in compressed sensing. *Signal Processing Letters, IEEE*, 18:478–481, 2011.
- [4] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.
- [5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.
- [6] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 2nd edition, 1985.
- [7] JW Boardman. Spectral Angle Mapping: A Rapid Measure of Spectral Similarity, 1993.
- [8] D. Brady and M. Gehm. Compressive imaging spectrometers using coded apertures. In *Proc. of SPIE*, volume 6246, pages 62460A–1 – 62460A–9, 2006.
- [9] E.J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406 – 5425, 2006.

- [10] C. Chang. Virtual dimensionality for hyperspectral imagery. *SPIE Newsroom*, 10(2.1200909):1749, 2009.
- [11] C. Chang and Q. Du. Estimation of number of spectrally distinct signal sources in hyperspectral imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(3):608 – 619, march 2004.
- [12] M. Davenport, M. Duarte, M. Wakin, J. Laska, D. Takhar, K. Kelly, and R. Baraniuk. The smashed filter for compressive classification and target recognition. In *Proceedings of SPIE*, volume 6498, 2007.
- [13] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces*, volume I, pages 317–366. North-Holland, Amsterdam, 2001.
- [14] Richard A. DeVerse, Ronald R. Coifman, Andreas C. Coppi, William G. Fateley, Frank Geshwind, Robert M. Hammaker, Sam Valenti, Fred J. Warner, and Gus L. Davis. Application of spatial light modulators for new modalities in spectrometry and imaging. In Richard M. Levenson, Gregory H. Bearman, and Anita Mahadevan-Jansen, editors, *Spectral Imaging: Instrumentation, Applications, and Analysis II*, volume 4959, pages 12–22. SPIE, 2003.
- [15] D. Donoho. Compressed sensing. *IEEE Trans. Info. Th.*, 52(4):1289–1306, 2006.
- [16] M.F. Duarte, M.A. Davenport, M.B. Wakin, and R.G. Baraniuk. Sparse signal detection from incoherent projections. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, 2006.
- [17] J. Fowler and Q. Du. Anomaly detection and reconstruction from random projections. *IEEE Transactions on Image Processing*, 21(1):184 –195, Jan. 2012.
- [18] M. Gehm, R. John, D. Brady, R. Willett, and T. Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Opt. Express*, 15(21):14013–14027, 2007.
- [19] Z. Guo and S. Osher. Template Matching via L1 Minimization and Its Application to Hyperspectral Data. Accepted to Inverse Problems and Imaging (IPI), 2009.
- [20] J. Han and B. Bhanu. Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 40(6):1771–1784, 2007.
- [21] J. Haupt, R. Castro, R. Nowak, G. Fudge, and A. Yeh. Compressive sampling for signal classification. In *Fortieth Asilomar Conference on Signals, Systems and Computers*, pages 1430–1434, 2006.

- [22] G. Healey and D. Slater. Models and methods for automated material identification in hyperspectral imagery acquired under unknown illumination and atmospheric conditions. *IEEE Transactions on Geoscience and Remote Sensing*, 37(6), 1999.
- [23] Alfred O Hero. Geometric entropy minimization (GEM) for anomaly detection and localization. In *In Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 585–592. MIT Press, 2006.
- [24] X. Jin, S. Paswaters, and H. Cline. A comparative study of target detection algorithms for hyperspectral imagery. In *Proceedings of SPIE*, volume 7334, page 73341W, 2009.
- [25] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. *Contemp. Math.*, pages 189–206, 1984.
- [26] W. Johnson, D. Wilson, W. Fink, M. Humayun, and G. Bearman. Snapshot hyperspectral imaging in ophthalmology. *Journal of Biomedical Optics*, 12(1):014036–1 – 014036–7, 2007.
- [27] E. Kelly. An adaptive detection algorithm. *IEEE Transactions on Aerospace and Electronic Systems*, pages 115–127, 1986.
- [28] F. Krahmer and R. Ward. New and improved johnson-lindenstrauss embeddings via the restricted isometry property. *Arxiv preprint arXiv:1009.0744*, 2010.
- [29] S. Kraut, L. Scharf, and L. McWhorter. Adaptive subspace detectors. *Signal Processing, IEEE Transactions on*, 49(1):1–16, Jan 2001.
- [30] K. Krishnamurthy, M. Raginsky, and R. Willett. Hyperspectral target detection from incoherent projections. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 3550–3553, 2010.
- [31] K. Krishnamurthy, M. Raginsky, and R. Willett. Hyperspectral target detection from incoherent projections: nonequiprobable targets and inhomogenous SNR. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 1357–1360, 2010.
- [32] FA Kruse, JW Boardman, AB Lefkoff, JM Young, KS Kierein-Young, TD Cocks, R. Jensen, and PA Cocks. HyMap: an Australian hyperspectral sensor solving global problems-results from USA HyMap data acquisitions. In *Proc. of the 10th Australasian Remote Sensing and Photogrammetry Conference*, 2000.
- [33] H. Kwon and N. Nasrabadi. Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 43(2):388–397, Feb. 2005.

- [34] H. Kwon and N. Nasrabadi. Kernel matched subspace detectors for hyperspectral target detection. *IEEE transactions on pattern analysis and machine intelligence*, pages 178–194, 2006.
- [35] R. Lin, B. Dennis, and A. Benz. *The Reuven Ramaty High-Energy Solar Spectroscopic Imager (RHESSI) - Mission Description and Early Results*. Kluwer Academic Publishers, Dordrecht, 2003.
- [36] D. Manolakis, D. Marden, and G. Shaw. Hyperspectral image processing for automatic target detection applications. *Lincoln Laboratory Journal*, 14(1):79–116, 2003.
- [37] D. Manolakis and G. Shaw. Detection algorithms for hyperspectral imaging applications. *IEEE Signal Processing Magazine*, 19(1):29–43, 2002.
- [38] M. Martin, S. Newman, J. Aber, and R. Congalton. Determining forest species composition using high spectral resolution remote sensing data. *Remote Sensing of Environment*, 65(3):249–254, 1998.
- [39] M. Martin, M. Wabuyele, K. Chen, P. Kasili, M. Panjehpour, M. Phan, B. Overholt, G. Cunningham, D. Wilson, R. DeNovo, and T. Vo-Dinh. Development of an advanced hyperspectral imaging (HSI) system with applications for cancer detection. *Annals of biomedical engineering*, 34(6):1061–1068, 2006.
- [40] J. Miller, C. Elvidge, B. Rock, and J. Freemantle. An airborne perspective on vegetation phenology from the analysis of AVIRIS data sets over the Jasper ridge biological preserve. In *Geoscience and Remote Sensing Symposium (IGARSS'90): Remote sensing for the nineties*, pages 565–568, 20-24 May 1990.
- [41] M. Parmar, S. Lansel, and B. Wandell. Spatio-spectral reconstruction of the multispectral datacube using sparse recovery. In *15th IEEE International Conference on Image Processing*, pages 473–476, 2008.
- [42] I. Reed and X. Yu. Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(10):1760–1770, Oct 1990.
- [43] L. Scharf and B. Friedlander. Matched subspace detectors. *IEEE Transactions on Signal Processing*, 42(8):2146–2157, 1994.
- [44] L.L. Scharf and L.T. McWhorter. Adaptive matched subspace detectors and adaptive coherence estimators. In *Conference Record of the Thirtieth Asilomar Conference on Signals, Systems and Computers*, pages 1114–1117, 1996.
- [45] D. Stein, S. Beaven, L. Hoff, E. Winter, A. Schaum, and A. Stocker. Anomaly detection from hyperspectral imagery. *IEEE Signal Processing Magazine*, 19(1):58–69, 2002.

- [46] Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *J. Machine Learning Research*, 6:211–232, 2005.
- [47] C. Stellan, G. Hazel, F. Bucholtz, J. Michalowicz, A. Stocker, and W. Schaaf. Real-time hyperspectral detection and cuing. *Optical Engineering*, 39, 2000.
- [48] J. Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, pages 2013–2035, 2003.
- [49] A. Szlam, Z. Guo, and S. Osher. A split Bregman method for non-negative sparsity penalized least squares with applications to hyperspectral demixing. In *IEEE 17th International Conference on Image Processing (ICIP)*, pages 1917–1920, 2010.
- [50] D. Takhar, J. Laska, M.B. Wakin, M.F. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R.G. Baraniuk. A new compressive imaging camera architecture using optical-domain compression. In *Proc. IS&T/SPIE Symposium on Electronic Imaging*, 2006.
- [51] T. Tao. Talagrand’s concentration inequality. <http://terrytao.wordpress.com/2009/06/09/talagrand-concentration-inequality/>. Accessed: 08/03/2012.
- [52] T. Tao and V. Vu. On random ± 1 matrices: singularity and determinant. *Random Structures & Algorithms*, 28(1):1–23, 2006.
- [53] A. Wagadarikar, R. John, R. Willett, and D. Brady. Single disperser design for coded aperture snapshot spectral imaging. *Appl. Opt.*, 47(10):B44–B51, 2008.
- [54] L. Wasserman. *All of Statistics*. Springer, 2004.
- [55] G. Wei, L. Agnihotri, and N. Dimitrova. TV program classification based on face and text processing. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 3, pages 1345 –1348 vol.3, 2000.
- [56] R. Willett, M. Gehm, and D. Brady. Multiscale reconstruction for computational spectral imaging. *Computational Imaging V*, 6498:64980L–1 – 64980L–15, 2007.
- [57] F. Woolfe, M. Maggioni, G. Davis, F. Warner, R. Coifman, and S. Zucker. Hyper-spectral microscopic discrimination between normal and cancerous colon biopsies. Manuscript, 2006.
- [58] K. Zuzak, S. Naik, G. Alexandrakis, D. Hawkins, K. Behbehani, and E. Livingston. Intraoperative bile duct visualization using near-infrared hyperspectral video imaging. *The American Journal of Surgery*, 195(4):491–497, 2008.

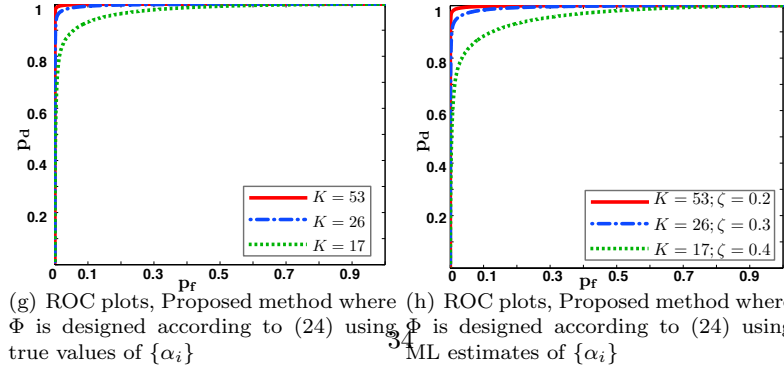
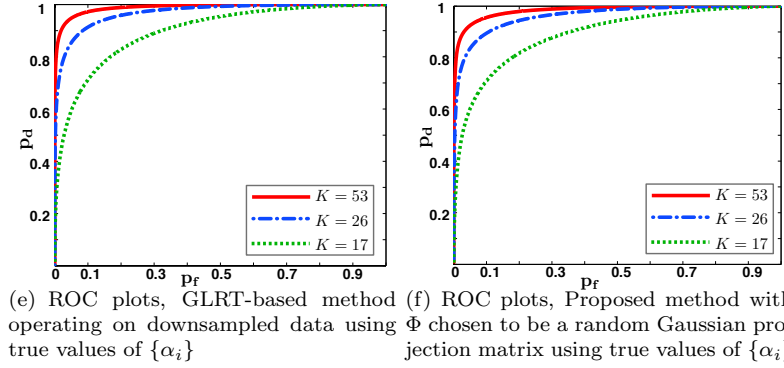
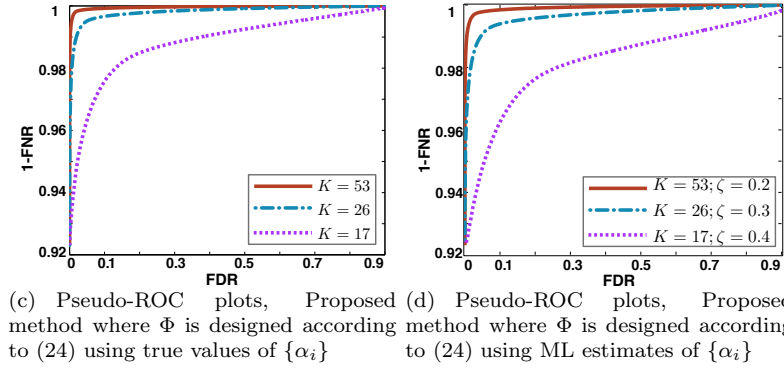
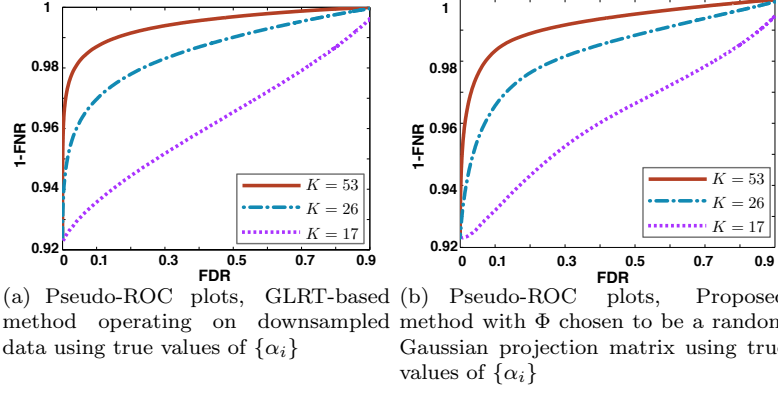


Figure 2: Comparison of the performances of the proposed anomaly detector using a random Φ , the proposed anomaly detector using the designed Φ in (24) and the GLRT-based method operating on downsampled data for different values of K when $\alpha_i^* \in \mathcal{U}[2, 3]$ and $\alpha_i = \alpha_i^* \sqrt{K}$.

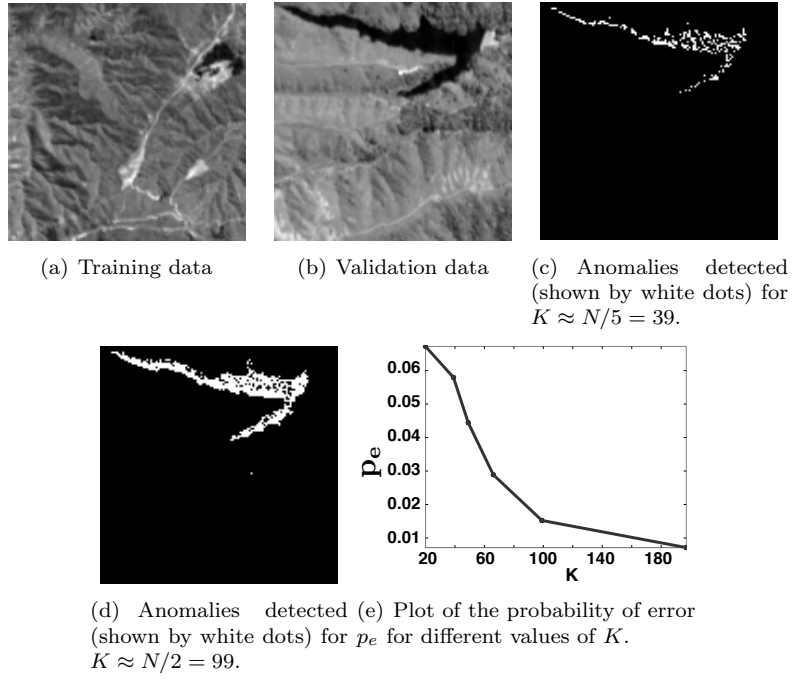


Figure 3: Anomaly detection results corresponding to real AVIRIS data for a fixed FDR control of 0.01.